STATISTICAL METHODS FOR THE ANALYSIS OF DESIGNS INCLUDING TREATMENTS DELIVERED TO GROUPS AND TO INDIVIDUALS: AN ANALYTIC AND MONTE CARLO STUDY

By

STEPHANIE BILLER WEHRY

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2001

ACKNOWLEDGEMENTS

I would like to thank my dissertation committee chairman, committee members, the faculty and staff of the School of Teaching and Learning, Dr. Adel Boules, of the University of North Florida and my family for their support and encouragement in my pursuit of this degree.

First, I thank my dissertation chairperson, Dr. James J. Algina, for suggesting the topic, guiding me through the process, and not giving up when the topic appeared to prematurely terminate. Additionally, I thank Dr. Algina for caring enough about the art of teaching to approach each lesson and assessment opportunity in a thoughtful, well-prepared manner--the mark of an excellent teacher.

Likewise, I also thank dissertation committee members Dr. Linda Crocker and Dr. David Miller for carefully and thoughtfully reading the manuscript, for excellent classes, and for the financial support of graduate assistantships. Furthermore, I thank committee member Dr. Donald Bernard for his support of my degree program from my initial visit with him in the summer of 1996 through the completion of the program.

Dr. Bernard, Dr. Mary G. Kantowski, and Dr. Ben Nelms have also provided financial support through graduate teaching assistantships in the School of Teaching and Learning. Faculty members and staff in the School of Teaching and Learning supported my teaching efforts and provided opportunities for me to continue interacting with pre-service elementary school teachers and with professional teaching and educational research organizations.

I acknowledge tremendous gratitude to Dr. Adel Boules of the University of North Florida Department of Mathematics and Statistics. I thank him for his assistance with the numerical methods necessary for the completion of this research.

Finally I thank my family for their patience and support.

TABLE OF CONTENTS

CHAPTERS

Abstract of Dissertation Presented to the Graduate School of the University of Florida in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

STATISTICAL METHODS FOR THE ANALYSIS OF DESIGNS INCLUDING TREATMENT DELIVERED TO GROUPS AND TO INDIVIDUALS: AN ANALYTIC AND MONTED CARLO STUDY

By

Stephanie Biller Wehry

May, 2001

Chair: James J. Algina
Major Department: Educational Psychology

Comparing treatment effectiveness in groups versus individual research designs, designs in which treatment delivered to individuals is compared to treatment delivered to individuals nested in groups, is common in social science research. However, it has been shown the widespread practice of ignoring any possible group dependency in the data can result in hypothesis testing using an inflated Type I error rate. A quasi-F test statistic with a synthetic mean square error has been suggested for use in groups versus individual designs having equal sized groups nested in the group treatment level. The chi square distribution of the error term has degrees of freedom approximated using a two-moment approach.

In this study, the quasi-F test statistic was extended to include a four-moment and an arithmetic average of the two- and four-moment approaches for approximating the degrees of freedom of the synthetic error term. A quasi-F test statistic was also developed to include designs that have unequal sized groups. The Type I error rates of the resulting quasi-F tests were studied using both Monte Carlo simulations and analytic methods. The Factors considered in the study were the size and number of the groups, the degree of group dependency, the degree of homoscedasticity of the treatment level variances, and the rate of subject attrition.

The results of the study suggested for designs involving two groups nested in the group treatment level, none of the approaches to approximating the error term degrees of freedom resulted in tests that control the Type I error rate. Increasing the size of the two groups further inflated the Type I error rate. In designs involving three groups, the average-moment quasi-F test is recommended, and, in designs involving four or more groups, the either two-moment or average-moment quasi-F test is recommended. Subject attrition rate had no significant effect on the Type I error rate of any of the quasi-F tests in the conditions studied.

CHAPTER ONE
INTRODUCTION

Researchers in the social sciences often use research designs that compare the means of measurements obtained from individuals nested in groups to those of individuals not nested in groups. The groups versus individuals research design can arise from efficacy studies that use waitlist or null treatment control groups as well as from comparative studies having the specific intent to compare the effectiveness of treatments delivered to individuals nested in groups to those delivered to individuals. Researchers studying the efficacy of treatments delivered to small groups often use waitlist control groups (Burlingame, Kircher, & Taylor, 1994). In fact, the most common experimental research design in psychotherapy involves the use of a randomly assigned control condition which can feature a variety of no-treatment control schemes (Clarke, 1998). In these designs, no treatment is delivered to the control individuals; therefore, individual responses are independent. In particular, the use of waitlist control groups is ethically pleasing because control group participants receive treatment after some specified period of time. In comparative studies, when the measured outcomes of treatments delivered to individuals nested in groups are compared to outcomes of treatments delivered to individuals, the result is a groups versus individuals design. Theory might suggest group interaction enhances, inhibits, or has no effect on the treatment.

Some recent examples of research that used the groups versus individuals research design are by Webb (1999), by Boling and Robinson (1999), and by Bates, Thompson, and Flanagan (1999). In her unpublished dissertation, Webb studied the effects of a group counseling intervention on high school students diagnosed with Attention-Deficit Hyperactivity Disorder. Theory in counselor education suggested group intervention would be effective; therefore, the study compared the outcome measurements of individuals nested in small groups to those of the individuals in a null treatment control group. The use of a null treatment control group resulted in a groups versus individuals research design. Boling and Robinson investigated the effects of study environment on a measure of knowledge following a distance learning lecture. The three levels of study environment included a printed study guide accessed by individuals, an interactive multi-media study guide accessed by individuals, and a printed study guide accessed by cooperative study groups. This investigation is an example of research specifically designed to compare the group and individual environments. In the research, the same study guide was tailored to accommodate each environment. Bates et al. compared the effectiveness of a mood induction procedure administered to groups to the effectiveness of the same procedure administered to individuals. Theory suggested delivering the procedure to groups should be more efficient and provide more homogeneity in the administration of the procedure. This study exemplifies research designed specifically to compare identical treatments delivered to individuals and to individuals nested in groups. Even though theory suggested subjects participating in groups would likely be inhibited, and consequently the mood induction procedure not as effective, the researchers hoped to find no difference in the effectiveness of the procedures.

Research in group psychotherapy also utilizes the groups versus individuals research design. Burlingame, Kircher, and Taylor (1994) reported that seven meta-analyses of group therapy conducted in the 1980s compared the effectiveness of group to individually delivered therapy. Some studies in the meta-analyses used treatments specifically designed for individual treatment with both individuals and individuals nested in groups in order to compare treatment effectiveness and to endorse the cost benefit and efficiency of group delivery. Other studies used treatments specifically designed to capitalize on the theorized effects of group behavior. In summary, both historical and current reviews of research support the use and continued use of groups versus individuals research designs in both efficacy and comparative studies of treatment effectiveness.

Methodological Concerns

Cook and Campbell (1979) used the term "statistical conclusion validity" to express concerns about the validity of conclusions formed from statistical relationships established through data analysis. Procedures that falsely hide or inflate statistical relationships raise issues of statistical conclusion validity. Low statistical validity can result from either using research designs that have inadequate statistical power or using designs that violate the assumptions of the mathematical models. Violating the assumption of independent observations across experimental units is a major concern in groups versus individuals research designs.

The mathematics of the F statistic shows the bias of the F statistic when observations from experimental units are dependent. Using the notation of Myers, Dicecco, and Lorch (1981), for a design having two treatment levels with g groups nested in the group treatment level and n subjects in each group, there are $N_G = g{\cdot}n$ subjects in the group treatment level and N_I subjects in the individual treatment level. The score of the i^{th} subject in the individual treatment level is Y_{iI}, and the score of the i^{th} subject in the j^{th} group in the group treatment level is Y_{ijG}. Both Y_{iI} and Y_{ijG} are composites of μ_I or μ_G, the population means of the treatment levels, and ε_{iT_I} or ε_{ijT_G}, the deviations of the observed score from its population mean: $Y_{iI} = \mu_I + \varepsilon_{iT_I}$ and $Y_{ijG} = \mu_G + \varepsilon_{ijT_G}$. Both ε_{iT_I} and ε_{ijT_G} have expected values of zero and variances of σ^2_{eI} and σ^2_G, respectively. However $E\left(\varepsilon_{ijT_G}, \varepsilon_{i'jT_G}\right) = \rho\sigma^2_G \neq 0$, where ρ is the correlation of pairs of scores within a single group in the group treatment level.

In order to have a central F distribution, the mean squares of the F ratio must estimate the same combination of population variances when the null hypothesis is true. Table 1 shows the sources of variance, degrees of freedom, mean squares, and expected mean squares for a completely balanced groups versus individuals design in which, for the sake of simplicity, it is assumed $\sigma^2_{eI} = \sigma^2_G = \sigma^2$. When the effect of treatment is tested against a pooled residual mean square, the result is a biased F statistic. The residual mean square is formed as a weighted composite of the three sources of variances. This pooling results in

$$MS_{residual} = \frac{(gn-1)\left(MS_{S/T_I}\right) + (g-1)\left(MS_{G/T_G}\right) + g(n-1)MS_{S/G/T_G}}{(gn-1) + (g-1) + g(n-1)}$$

$$\text{and } E(MS_{residual}) = \frac{2\sigma^2(gn-1)-(n-1)\rho\sigma^2}{2(gn-1)},$$

where MS_{G/T_G} is the mean square of the groups nested in the group treatment level, $MS_{S/G/T_G}$ is the mean square of the subjects nested in groups nested in the group treatment level, and MS_{S/T_I} is the mean square of subjects nested in the individual treatment level. Even when H_O is true, the expected value of the mean square for treatment effectiveness is not equal to the expected value of the pooled residual term. As long as $\rho\sigma^2 \neq 0$, the F ratio will be biased. In most instances, the correlation between observations within a group will be positive resulting in a positive bias and an inflation of the Type I error rate.

Table 1

ANOVA Table for a Basic Groups Versus Individuals Balanced Research Design

SV	df	MS	EMS
Treatment	1	$\frac{gn}{2}(\bar{Y}_{.I}-\bar{Y}_{..G})^2$	$\sigma^2+\left(\frac{n-1}{2}\right)\rho\sigma^2+\frac{gn}{2}(\mu_I-\mu_G)^2$
Subjects within Individual Treatment Level	*gn*-1	$\sum_{i=1}^{gn}(Y_{iI}-\bar{Y}_{.I})^2/(gn-1)$	σ^2
Groups within Group Treatment Level	*g*-1	$n\sum_{j=1}^{g}(\bar{Y}_{.jG}-\bar{Y}_{..G})^2/(g-1)$	$\sigma^2+(n-1)\rho\sigma^2$
Subjects within the Group Treatment Level	*gn*-*g*	$\sum_{j=1}^{g}\sum_{i=1}^{n}(Y_{ijG}-\bar{Y}_{.jG})^2/(gn-g)$	$\sigma^2-\rho\sigma^2$
Total	2*gn*-1		

Myers et al. (1981), Kromrey and Dickinson (1996), and Burlingame, Kircher, and Honts (1994) conducted Monte Carlo studies of the Type I error rates of research designs having groups nested in treatments when the analysis ignores the group effect. The Type I error rates of simulated data for designs having two treatment levels and with the number of subjects in the treatment levels, the number of groups, the number of subjects in the groups, and the levels of dependency manipulated are presented in Table 2. In all cases, increasing the dependency of the data resulted in increased Type I error rates. The use of waitlist control groups in the Burlingame et al. study indicated the Type I error rate for this design is lower than for designs in which both treatment levels use subjects nested in groups. However, the Type I error rates for a nominal alpha of 0.05 are 0.18 and 0.23 for intraclass correlations of 0.20 and 0.40, respectively. The Kromrey and Dickinson study indicated that increasing the number of groups made little difference in the inflated Type I error rate but increasing the number of individuals within the groups even further inflated the Type I error rates. Over all conditions of the studies, Type I error rates ranged from a low of 0.066 to a high of 0.607 when the nominal alpha level was 0.05.

Unfortunately, biased F statistics and the resulting inflated Type I error rates are routinely reported in studies because of the common practice of analyzing groups-nested-in-treatment data at the individual observations level. Burlingame, Kircher, and Taylor (1994) reported, in a review of group psychotherapy literature published from 1980 to 1992, most experimental research used t-tests, ANOVA, or ANCOVA. Even though these statistical procedures require independent observations across experimental units, 89% of the studies treated observations from individuals nested in groups as though they were statistically independent; furthermore, researchers made no attempt to address any degree of dependency that might exist.

Table 2

Mean Type I Error Rates for a Balanced, Two-Treatment Level, Groups Versus Individuals Design When Data Are Analyzed at the Individual Level Using a Pooled Residual Error Term: Nominal Alpha = 0.05

Study	*g*	*n*	Type I Error Rate: ρ_{ICC} = 0.20	Type I Error Rate: ρ_{ICC} = 0.40	Type I Error Rate: ρ_{ICC} = 0.80
Myers et al.	2 (waitlist)	4		0.076	0.103
	2 (waitlist)	8		0.066	0.098
	5 (waitlist)	4		0.144	0.243
	5 (waitlist)	8		0.158	0.237
Burlingame et al.	2	8	0.225	0.372	
	2 (waitlist)	8	0.181	0.230	
Kromrey &	2	3	0.100	0.152	
Dickinson	2	10	0.251	0.340	
	2	30	0.466	0.607	
	3	3	0.103	0.150	
	3	10	0.249	0.395	
	3	30	0.470	0.600	
	5	3	0.096	0.155	
	5	10	0.249	0.385	
	5	30	0.470	0.593	

Following the same trend, the studies by Webb (1999), Boling and Robinson (1999), and Bates et al. (1999) made no reference to the dependency of observations collected from individuals nested in groups. In each of these three studies, the unit of analysis was individual observations. As shown in cases of dependency, tests of treatment effectiveness are conducted with an inappropriate pooled error term and its associated degrees of freedom. However, when group means rather than the individual observations are used as the unit of analysis, more data may be required to achieve the same level of statistical power.

At least three avenues of methodological research have addressed the dependency of data collected from individuals nested in groups. One avenue is the development of statistical methods to adequately test any groups-nested-in-treatment effect in the data before making the decision to ignore the group effect and conduct the analysis at the individual observations level. This avenue attempts to maintain the advantage of the statistical power enjoyed by analysis at the individual level. A second avenue is the development of nonparametric tests as alternatives to the t-test, ANOVA, and ANCOVA. One such example is the use of bootstrapping. A third avenue is the development of random effects ANOVA models that accommodate both the group and individual data through the use of quasi-F test statistics and pseudogroup procedures.

Testing the Dependency of Group Data

Hopkins (1982) suggested, in comparative group studies with completely balanced designs (designs having equal numbers of subjects nested in an equal number of groups across all treatment levels), using a random effects ANOVA model in order to test the null hypothesis of no groups-nested-in-treatment effect. The model for this approach is

$$y_{ijk} = \mu + \alpha_k + \beta_{j:k} + \varepsilon_{i:jk},$$

where y_{ijk} is the observation of the i^{th} subject ($i = 1,\ldots, n$) nested in the j^{th} group ($j = 1,\ldots, g$) and in the k^{th} treatment ($k = 1,\ldots, t$). In the model, μ is the grand mean, α_k is the fixed effect of the k^{th} treatment, $\beta_{j:k}$ is the random effect of the j^{th} group nested in the k^{th} treatment, and $\varepsilon_{i:jk}$ is the residual of the i^{th} subject nested in the j^{th} group and in the k^{th} treatment. It is assumed that $\beta_{j:k}$ and $\varepsilon_{i:jk}$ have means of zero and variances of σ_g^2 and σ_ε^2, respectively. The null hypothesis of no groups-nested-in-treatment effect can be stated as

$$H_O{:}\sigma_g^2 = 0.$$

If the test of this hypothesis yields a statistically nonsignificant result, the group effect is eliminated from the model, and the group data are analyzed at the individual observation level. If the test yields a statistically significant result, the data are analyzed at the group level using group means as the experimental unit of analysis. Because ignoring the group effect results in a pooled error term, Hopkins suggested the null hypothesis of no groups-nested-in-treatment effect be tested with an adjusted nominal alpha level of 0.20 to 0.25.

Kromrey and Dickinson (1996) conducted a Monte Carlo study of Hopkins' suggestion in which they manipulated the number of fixed treatments, the number of random groups nested in treatment levels, the number of subjects in the groups, the degree of dependency among observations within a group, and the effect size. The results of the study, presented in Table 3, indicated the power for the test of $H_O:\sigma_g^2= 0$ is greater for more groups nested in the treatments and for more subjects nested in each group. The power of the test also increased as the dependency among observations within the groups increased; however, as this dependency increases so does the bias of the F statistic.

Table 3

Power of ANOVA Tests of Groups Within Treatments Effects for Two Treatment Levels, α= 0. 05 and 0.30, and Three Levels of Intraclass Correlation

g	n	α	$\rho_{ICC} = 0.00$	$\rho_{ICC} = 0\ .20$	$\rho_{ICC} = 0.40$
2	3	0.05	0.051	0.145	0.283
2	3	0.30	0.297	0.482	0.635
2	10	0.05	0.058	0.397	0.663
2	10	0.30	0.561	0.697	0.860
2	30	0.05	0.047	0.714	0.856
2	30	0.30	0.294	0.879	0.939
3	3	0.05	0.054	0.194	0.412
3	3	0.30	0.314	0.564	0.763
3	10	0.05	0.048	0.573	0.849
3	10	0.30	0.294	0.840	0.951
3	30	0.05	0.051	0.886	0.974
3	30	0.30	0.296	0.965	0.994

Note: Power estimates are based on 5000 samples of each condition.

Unfortunately for researchers wishing to ignore group dependency, Kromrey and Dickinson also found the bias of the F test of treatment effectiveness using individual observations as the unit of analysis increased more rapidly than the power of the test to detect the dependency of the observations. Increasing nominal alpha to 0.30 in order to increase the power of the test of $H_O\!:\sigma_g^2 = 0$ greatly reduced subsequent Type I error rates when data were analyzed at the individual observation level with a nominal alpha of 0.05. However, a nominal alpha of 0.30 was not sufficient for all conditions studied and did not control Type I error rates as well with few groups or with few subjects in the groups. Kromrey and Dickinson investigated only comparative, completely balanced designs in which all levels of the treatment had equal number subjects in equal number of groups.

Nonparametric Bootstrapping

Burlingame, Kircher, and Honts (1994) suggested using bootstrapping as an alternative to the random effects ANOVA model when the groups-nested-in-treatment effect is significant at an adjusted nominal alpha level. Bootstrapping is a nonparametric procedure in which the sampling distribution of the test statistic is empirically determined through resampling. Bootstrapping has been shown to be a useful alternative to parametric test statistics when assumptions underlying the mathematical model of the parametric test statistics are violated. Burlingame, Kircher, and Honts conducted research to investigate the use of bootstrapping when data are dependent across experimental units. In the Monte Carlo study, Burlingame, Kircher, and Honts manipulated the number of treatments and groups, the degree of dependency, mixed levels of dependency in order to simulate waitlist control groups, and effect size. Data were analyzed at the individual observations level in order to simulate the common research practice of ignoring any group effect. Eight subjects were simulated in all groups based on the review of psychotherapy literature in which eight was found to be the median number of individuals per group (Burlingame, Kircher, & Taylor, 1994).

The results of the simulation, reported in Table 4, suggested bootstrap was generally less sensitive to dependency of observations than ANOVA. In cases of positive dependency, both methods resulted in inflated Type I error rates. The use of a waitlist control group resulted in both ANOVA and bootstrap methods maintaining Type I error rates closer to the nominal alpha with bootstrap exhibiting better control across the conditions of the study. The power of both the ANOVA and bootstrap were similar: averaged across all conditions studied, 31.1% for the ANOVA test and 31.5% for the bootstrap. Burlingame, Kircher, and Honts (1994) suggested researchers first conduct ANOVA tests at both the individual and group levels. If these tests yield conflicting results, use bootstrap to decide the effectiveness of the treatment. Similar to the simulation of Kromrey and Dickinson (1996), Burlingame, Kircher, and Honts simulated a completely balanced design in which all levels of the treatment had equal numbers of subjects in an equal number of groups.

Table 4

The Type I Error Rates of ANOVA and Bootstrapping in Conditions Involving Two Treatment Levels: Nominal Alpha = 0. 05

g	n	Test	$\rho_{ICC} = 0.00$	$\rho_{ICC} = 0.20$	$\rho_{ICC} = 0.40$
2	8	ANOVA	0.052	0.255	0.375
2	8	Bootstrap	0.056	0.109	0.123
3	8	ANOVA	0.049	0.257	0.375
3	8	Bootstrap	0.050	0.107	0.127
2 (Waitlist)	8	ANOVA	0.049	0.181	0.230
2 (Waitlist)	8	Bootstrap	0.051	0.068	0.067

Note: Results are based on 1000 samples of each condition.

Quasi-F and Pseudo-F Tests

Pseudogroup tests

If the group treatment level has g groups with n subjects in each group, and the individual treatment level contains N_I individuals, where N_I is a multiple of g; the N_I subjects can be randomly placed, after treatment, into g groups having n_i subjects in each group. The observations from subjects within these pseudogroups remain independent, but an unbiased F test, or one-way ANOVA, of the treatment effect can be conducted using the group and pseudogroup means as units of the analysis (Myers et al., 1981). The model violates the assumption of homogeneity of variance across treatments, but heterogeneity of variance does not usually result in a large distortion of the Type I error rate in balanced data provided the heterogeneity is not too large

(Wilcox, 1988). If the number of groups across treatment levels is not the same or if the heterogeneity is large, the F test is biased and the Type I error rate may be distorted.

Quasi-F tests

The major problem in groups versus individuals designs is testing the effectiveness of the treatment with an appropriate error term. The error term must have the same expected value as the treatment effect. Myers et al. (1981) proposed using an error term formed as a linear combination of mean squares for the groups versus individuals research design having two levels of treatment with a possible unequal number of subjects in each treatment level but with the same number of subjects nested in each group. This synthetic error term for the test of treatment effectiveness is

$$MS_e = \frac{N_I}{N} MS_{G/T_G} + \frac{N_G}{N} MS_{S/T_I} \qquad (1)$$

where N_G is the number of subjects in the group treatment level and is equal to $n \cdot g$, N is the total number of participants and is equal to $N_G + N_I$. The mean square MS_{G/T_G} is

$$n\sum_{j=1}^{g}\left(\bar{Y}_{.jG} - \bar{Y}_{..G}\right)^2 / (g-1),$$

and MS_{S/T_I} is

$$\sum_{i=1}^{N_I}\left(Y_{iI} - \bar{Y}_{.I}\right)^2 / (N_I - 1),$$

where $\bar{Y}_{jG}$ is the mean of the j^{th} group in the group treatment level, $\bar{Y}_{..G}$ is the mean of the group treatment level, Y_{iI} is the observation of the i^{th} subject in the individual treatment level, and $\bar{Y}_{.I}$ is the mean of the individual treatment level. The numerator for the quasi-F test is

$$\frac{\left(\bar{Y}_{.I}-\bar{Y}_{..G}\right)^2}{\left(\frac{1}{N_I}+\frac{1}{N_G}\right)}.$$

The degrees of freedom for a synthetic mean square can be approximated using Satterthwaithe's two-moment approximation (Satterthwaite, 1941). If the synthetic mean square is formed as a linear combination such that $MS = a_1 MS_1 + \ldots + a_k MS_k$, the associated approximate degrees of freedom are

$$\hat{df} = \frac{(MS)^2}{\frac{(a_1 MS_1)^2}{df_1}+\cdots+\frac{(a_k MS_k)^2}{df_k}}.$$

The degrees of freedom for the synthetic mean square of the quasi-F test are

$$\hat{df} = \frac{\left(\frac{N_I}{N}MS_{G/T_G}+\frac{N_G}{N}MS_{S/T_I}\right)^2}{\frac{\left(\frac{N_I}{N}MS_{G/T_G}\right)^2}{g-1}+\frac{\left(\frac{N_G}{N}MS_{S/T_I}\right)^2}{N_I-1}}. \tag{2}$$

The quasi-F procedure provides an unbiased error term with which to test treatment effectiveness.

Comparison of the pseudogroup and quasi-F procedures

Myers et al. (1981) conducted a Monte Carlo simulation to compare the use of quasi-F and pseudogroup procedures in tests of treatment effectiveness. They manipulated group dependency, group size and number, effect size, and the number of subjects in the individual level of the treatment. The Type I error rates, reported in

Table 5, suggested both measures reasonably control Type I error rates at the nominal level, and they reported the power of the two methods is about the same.

Table 5

Type I Error Rates for Pseudogroup and Quasi–F Tests of Treatment Effectiveness in Groups Versus Individuals Research Designs: Nominal Alpha=0.05

g	n	N_I	Test	ρ=0.40	ρ=0.80
2	4	4	Pseudogroup	0.0595	0.0520
2	4	4	Quasi-F	0.0505	0.0425
2	4	8	Pseudogroup	0.0500	0.0480
2	4	8	Quasi-F	0.0455	0.0495
2	8	8	Pseudogroup	0.0635	0.0590
2	8	8	Quasi-F	0.0580	0.0555
2	8	16	Pseudogroup	0.0525	0.0490
2	8	16	Quasi-F	0.0465	0.0435
5	4	4	Pseudogroup	0.0580	0.0460
5	4	4	Quasi-F	0.0485	0.0355
5	4	20	Pseudogroup	0.0480	0.0590
5	4	20	Quasi-F	0.0520	0.0615
5	8	8	Pseudogroup	0.0555	0.0480
5	8	8	Quasi-F	0.0545	0.0465
5	8	40	Pseudogroup	0.0590	0.0620
5	8	40	Quasi-F	0.0535	0.0550

Note: Results are based on 2000 samples of each condition.

In a few cases, the pseudogroup approach showed slightly better power, usually about 4%; therefore, Myers et al. (1981) suggested using the pseudogroup approach whenever possible. Increasing the number of subjects in the individual treatment level increased the power of both procedures. Because the number of groups figures directly in the calculations of the degrees of freedom for both tests, increasing the number of groups also had a large impact on the power of both tests.

Statement of the Problem

The variables manipulated in the three Monte Carlo studies (Burlingame, Kircher, & Honts, 1994; Kromrey & Dickinson, 1996; Myers et al., 1981) are summarized in Table 6. Only Burlingame, Kircher, and Honts and Myers et al. specifically studied conditions that give rise to the groups versus individuals research design. Because they did not vary dependency across treatment levels, Kromrey and Dickinson only investigated research designs having groups nested in all treatment levels. However, the test of no-groups-nested-in-treatment effect can be conducted using one-way, random effects ANOVA on the portion of data that arises from groups nested within the treatment levels.

All three Monte Carlo studies only simulated designs in which data were balanced across groups, resulting in the same number of subjects nested in each group. However, unequal sized groups can arise through the use of convenience samples, through the use of groups formed in context such as family units, and through attrition of research subjects. Researchers can balance designs across groups by randomly eliminating data in order to achieve equal sized groups or they can use methodology that accommodates unbalanced data. Eliminating data results in a loss of statistical power but methods that test treatment effectiveness in groups versus individuals research designs involving unbalanced data have not been studied in terms of Type I error rates and statistical power. One purpose of this study was to extend the quasi-F test statistic of Myers et al. to include designs that are not balanced across groups (unequal number of subjects nested in each group) as well as not balanced across treatment levels.

A second purpose of this study was to extend the work of Myers et al. (1981) to include new methods of approximating the degrees of freedom of the denominator synthetic mean square. The degrees of freedom used by Myers et al. are based on approximating the distribution of the mean square error of equation (1) by a multiple of a central chi-square distribution such that the mean and variance of the approximating distribution are equal to those of the mean square error. This approach is referred to as a two-moment approximation. The first step was to extend the Myers et al. study of balanced data by using a four-moment approximation of the degrees of freedom and an average of the two-moment and four-moment approximations. The effectiveness of the approach of Myers et al. and these new approximations in controlling the Type I error rate for balanced groups versus individuals designs was investigated using a Monte Carlo design. In addition, analytic results on the size of the tests resulting from the use of the three approaches for estimating the degrees of freedom were presented for balanced designs. Furthermore, the three approaches were extended for use in unbalanced groups versus individual designs, and the effectiveness of the approaches in controlling Type I error rates was investigated using a Monte Carlo design.

Table 6

Levels of Variables Manipulated in the Monte Carlo Studies

	Variables Manipulated								
Study	Treatments	g	ρ_{ICC}	Mixed Levels of Dependency	n_g	N_t	Completely Balanced	Balanced within Groups but not across Treatments	Effect Size
Myers et al.	2	4	0.00	yes	2	4	yes	yes	0.00
		3	0.40		5	8			0.25
			0.80			16			0.50
									0.75
Burlingame et al.	2	2	-0.05	yes	8	8	yes	no	0.00
	3	3	0.00						0.25
			0.20						0.50
			0.40						
Kromrey & Dickinson	2	2	0.00	no	3	NA	yes	no	0.00
	3	3	0.20		10				0.20
		5	0.40		30				0.50
			0.60						0.80
			0.80						
			1.00						

This research is important because groups versus individuals research designs are frequently used to investigate treatment effectiveness in the social sciences where unbalanced designs also frequently occur. Burlingame, Kircher, and Taylor (1994) reported in the 69% of comparative group studies that reported attrition rates, the rates ranged from 0% to 63% with a median value of 18% across the studies. Clarke (1998) further reported that the attrition rate among waitlist, null control treatment levels could even be higher than that of active treatment levels. In educational research, imbalance can occur because subjects nested in active treatment levels are also nested in unequally sized classrooms and because of attrition due to attendance and drop out rates of subjects.

CHAPTER TWO
METHODOLOGY

This chapter develops the quasi-F test statistic for comparing treatment effectiveness in groups versus individuals research designs having unbalanced data at both the treatment and group levels. The first section presents the statistical model for the test statistic used to compare two weighted means. However, to use the test statistic, one needs to estimate the variance components and to approximate the distribution of the synthetic error term. The second section discusses methods used to estimate variance components, and the following section presents the method of moments procedure estimates of the variance components. The fourth section provides the variance component estimates used to form the test statistic, and because the method of moments estimation procedure can result in negative estimates, the fifth section considers this issue. The sixth section discusses the second requirement for the use of the test statistic, methods used to approximate the distribution of the synthetic error term. The section presents three approaches for use with data balanced across groups and develops a modified method for use with data that are not balanced across groups. The seventh section discusses the variables that were manipulated in the Monte Carlo simulation and the analytic study investigating the Type I error rates of the three resulting quasi-F tests. The 5472 conditions of the study are summarized in the final section.

Quasi-F Test for Treatment Effectiveness Comparing Weighted Means

In the groups versus individuals research design, the linear model for the individual treatment level component is

$$Y_{iI} = \mu_I + \varepsilon_{iI} \text{, } i = 1,\ldots,N_I \text{,} \tag{3}$$

where Y_{iI} is the observation of the i^{th} subject in the individual treatment level, μ_I is the population mean of the individual treatment level, ε_{iI} is the residual of the i^{th} subject in the individual treatment level with $\varepsilon_{iI} \sim N\left(0,\sigma_{\varepsilon I}^2\right)$, and N_I is the total number of subjects in the individual treatment level. The linear model for the group treatment level component is

$$Y_{ijG} = \mu_G + \alpha_j + \varepsilon_{ijG} \text{,} \tag{4}$$

where Y_{ijG} is the observation of the i^{th} subject in the j^{th} group, $i = 1,\ldots,n_j$ and $j = 1,\ldots,g$, μ_G is the population mean of the group treatment level, α_j is the random effect of the j^{th} group with $\alpha_j \sim N\left(0,\sigma_g^2\right)$, and ε_{ijG} is the residual of the i^{th} subject in the j^{th} group in the group treatment level with $\varepsilon_{ijG} \sim N\left(0,\sigma_{\varepsilon G}^2\right)$. For a test of weighted means

$$\overline{Y}_{.I} = \frac{\sum_{i=1}^{N_I} Y_{iI}}{N_I} \tag{5}$$

and

$$\overline{Y}_{..G} = \frac{\sum_{j=1}^{g} n_j \overline{Y}_{.jG}}{N_G} \text{,} \tag{6}$$

where n_j is the number of subjects in the j^{th} group and $\bar{Y}_{.jG}$ is the mean of the j^{th} group in the group treatment level. Equation (5) is used to determine the variance of $\bar{Y}_{.I}$,

$$Var\left(\bar{Y}_{.I}\right) = \frac{\sigma^2_{eI}}{N_I}, \tag{7}$$

and equation (6) to determine the variance of $\bar{Y}_{..G}$,

$$Var\left(\bar{Y}_{..G}\right) = \frac{\sigma^2_g \sum_{j=1}^{g} n_j^2}{N_G^2} + \frac{\sigma^2_{eG}}{N_G}. \tag{8}$$

As is well known, if the variances of $\bar{Y}_{.I}$ and $\bar{Y}_{..G}$ were known, the hypothesis $H_0: \mu_I - \mu_G = 0$ could be tested by

$$\chi^2 = \frac{\left(Y_{.I} - Y_{..G}\right)^2}{Var\left(\bar{Y}_{.I} - \bar{Y}_{..G}\right)}. \tag{9}$$

Because observations are independent across treatment levels, substituting the variances from equations (7) and (8) into equation (9) results in

$$\chi^2 = \left(\bar{Y}_{.I} - \bar{Y}_{..G}\right)^2 / \left(\frac{\sigma^2_{eI}}{N_I} + \frac{\sigma^2_g \sum_{j=1}^{g} n_j^2}{N_G^2} + \frac{\sigma^2_{eG}}{N_G} \right). \tag{10}$$

However, the variances are not known, and, in order to develop a test statistic that can be used in practice, two steps must be completed: Develop estimators of the variance components in equation (10) and approximate the distribution of the resulting test statistic. Approximating the distribution of the denominator by a chi-square distribution and the distribution of the test statistic by an F distribution is a common practice in statistics.

Estimation of Variance Components – Weighted Means

There are numerous methods for estimating the variance components of an experimental design. The most commonly used method is the method of moments, also called the ANOVA estimation of variance components (Milliken & Johnson, 1992). The method of moments procedure is based on equating the expected values of partitioned sums of squares to their respective observed values. Within this method, one estimation procedure involves using the assumptions of the model to algebraically evaluate the expected mean squares. Other method of moments procedures, developed by Henderson (1953), use computer methods to find the expected values of the sum of squares. Yet another class of estimators involves maximum likelihood methods. Maximum likelihood (ML) estimators are values of the parameter space that maximize the likelihood function. Often, in data from balanced designs, the likelihood equations can be solved explicitly; however, for unbalanced designs, iterative methods are needed (Milliken & Johnson, 1992). In restricted maximum likelihood estimations (REML), the likelihood equations are partitioned into two parts, one part that is free of fixed effects. REML maximizes the part that has no fixed effects. Still a third class of estimators, described by Rao (1971), provides methods for obtaining minimum norm quadratic unbiased estimators (MINQUE) and minimum variance quadratic unbiased estimators (MIVQUE) of the variance components. These methods are iterative, and the researcher must provide initial values of the components. The estimates, therefore, depend on both the data and the initial values chosen. All methods produce the same results when the design is balanced (Milliken & Johnson, 1992; Swallow & Monahan, 1984).

Swallow and Monahan (1984) conducted a Monte Carlo study of ANOVA, ML, REML, MIVQUE and MINQUE methods of estimating the variance components of a one-way unbalanced, random effects design. All simulated data were normal, and the variables manipulated were the degree of imbalance, the number of groups, and the ratio of $\sigma_g^2/\sigma_{\varepsilon G}^2$. In terms of bias of the estimates, the results indicated, except in cases of extreme patterns of imbalance, $n_j = (1,1,1,1,13,\text{and } 13)$ and $n_j = (1,1,1,1,1,1,1,19,\text{and } 19)$, ANOVA, REML, and MINQUE estimators showed little difference. However, the results indicated that ML methods were the best estimators of σ_g^2 when $\sigma_g^2/\sigma_{\varepsilon G}^2 \leq .5$ because of the small bias and the low mean square error of the estimate. When $\sigma_g^2/\sigma_{\varepsilon G}^2$ is large, Swallow and Monahan indicated there may be a substantial downward bias and that ML methods have no superiority over the other methods. There was little difference among the methods studied when estimating $\sigma_{\varepsilon G}^2$. Milliken and Johnson (1992) suggested that ANOVA estimates should have good properties for nearly balanced data, and Swallow and Monahan concluded that unless the data are severely unbalanced and $\sigma_g^2/\sigma_{\varepsilon G}^2 > 1$, ANOVA estimates are adequate.

The results of the Swallow and Monahan (1984) study and the recommendations of Milliken and Johnson (1992) indicated, for the groups versus individuals research design, ANOVA estimates of the variance components are adequate. Data as extreme as that simulated in the Swallow and Monahan study would be rare in group research; therefore, method of moments estimators of the variance components are developed for the quasi-F test for comparing the effectiveness of two treatment levels when data are completely unbalanced.

Variance Component Estimates

Searle (1971) listed the assumptions of the linear model necessary for algebraically determining the expected values of the ANOVA sums of squares as the expected values of α_j, ε_{ijG}, and ε_{il} are all zero and the variances are σ_g^2, $\sigma_{\varepsilon G}^2$, and $\sigma_{\varepsilon l}^2$, respectively. Additionally, the α_js and the ε_{ijG}s are stochastically independent, that is,

$$E(\alpha_j, \alpha_{j'}) = 0 \quad \text{for } j \neq j', \text{ and}$$

$$E(\alpha_j, \varepsilon_{ij'G}) = 0 \quad \text{for all } i, j, \text{ and } j'. \quad \text{Additionally,}$$

$$Cov(Y_{ijG}, Y_{i'j'G}) = \begin{Bmatrix} \sigma_g^2 + \sigma_{\varepsilon G}^2 \text{ for } i = i' \text{ and } j = j' \\ \sigma_g^2 \text{ for} i \neq i' \text{ and } j = j' \\ 0 \text{ for} i \neq i' \text{ and } j \neq j' \end{Bmatrix}.$$

For the group treatment level variance components, the sums of squares are SS_{G/T_G}, the sum of the squared differences of the observed group means and the observed group treatment level mean, and $SS_{S/G/T_G}$, the sum of the squared differences of the individual observations and their observed group means. Similarly, for the individual treatment level variance component, the sum of squares, SS_{S/T_I}, is the sum of the squared differences of the individual observations and the observed individual treatment level mean.

The expected values for the corresponding mean squares are

$$EMS_{G/T_G} = \sigma_{\varepsilon G}^2 + n_o \sigma_g^2, \tag{11}$$

where

$$n_o = \frac{1}{g-1}\left(N_G - \frac{\sum_{j=1}^{g} n_j^2}{N_G} \right)$$

(Snedecor & Cochran, 1956). The other two expected values are

$$EMS_{S/G/T_G} = \sigma_{eG}^2 \tag{12}$$

and

$$EMS_{S/T_I} = \sigma_{eI}^2 . \tag{13}$$

Equations (11), (12), and (13) are solved for the ANOVA variance component estimates and the values substituted into equation (10) to obtain the quasi-F test statistic for comparing weighted treatment level means.

The Test Statistic

The resulting quasi-F test statistic is

$$\hat{F}_{quasi} = \frac{\left(\bar{Y}_{.I} - \bar{Y}_{..G}\right)^2}{\left\{ \frac{MS_{S/T_I}}{N_I} + \frac{\left(MS_{G/T_G} - MS_{S/G/T_G}\right)\sum_{j=1}^{g} n_j^2 / n_o}{N_G^2} + \frac{MS_{S/G/T_G}}{N_G} \right\}} , \tag{14}$$

which simplifies to

$$\hat{F}_{quasi} = \frac{\left(\bar{Y}_{.I} - \bar{Y}_{..G}\right)^2}{\left\{ \frac{MS_{S/T_I}}{N_I} + \frac{MS_{G/T_G}\sum_{j=1}^{g} n_j^2}{n_o N_G^2} + \frac{MS_{S/G/T_G}\left(n_o N_G - \sum_{j=1}^{g} n_j^2 \right)}{n_o N_G^2} \right\}} . \tag{15}$$

The denominator of the quasi-F statistic is a synthetic mean square in the form of $MS = a_1 MS_{S/T_I} + a_2 MS_{G/T_G} + a_3 MS_{S/G/T_G}$, where

$$a_1 = \frac{1}{N_I}, \tag{16}$$

$$a_2 = \frac{\sum_{j=1}^{g} n_j^2}{n_o N_G^2}, \tag{17}$$

and

$$a_3 = \left(\frac{n_o N_G - \sum_{j=1}^{g} n_j^2}{n_o N_G^2} \right). \tag{18}$$

Negative Variance Component Estimates

There is nothing in the method of moments variance component estimation procedure to prevent negative variance component estimates. Anytime $MS_{G/T_G} < MS_{S/G/T_G}$, $\hat{\sigma}_g^2$ is negative. One possible interpretation of a negative estimate is that σ_g^2 is actually zero in which case the group treatment level model reduces to $Y_{ijG} = \mu_G + \varepsilon_{ijG}$ (Searle, 1992). Searle also suggested researchers sometimes conceptualize the model covariance as

$$Cov\left(Y_{ijG}, Y_{i'j'G}\right) = \begin{Bmatrix} \sigma^2 \rho \text{ when } i \neq i' \text{ and } j = j' \\ \sigma^2 \text{ when } i = i' \text{ and } j = j' \\ 0 \text{ when } i \neq i' \text{ and } j \neq j' \end{Bmatrix},$$

where ρ is the correlation between Y_{ijG}s from the same group. In this model, negative estimates are allowable negative correlations, and, for balanced data,

$$\hat{\rho} = \left(MS_{G/T_G} - MS_{S/G/T_G}\right) \Big/ \left((n-1)MS_{S/G/T_G} + MS_{G/T_G}\right),$$

$$MS_{S/G/T_G} = \left(1 - \hat{\rho}\right)\hat{\sigma}^2, \text{ and}$$

$$MS_{G/T_G} = (n\hat{\rho} + 1 - \hat{\rho})\hat{\sigma}^2,$$

where n is the number of subjects nested in each group. Myers et al. (1981) used this approach in forming the linear model of the group treatment level component in their study. However, when the parameters ρ and σ^2 are used to solve for σ^2_g and σ^2_{gG}, the problem of negative variance component estimates returns.

For data that are balanced across groups, the probability that $\hat{\sigma}^2_g$ is negative is the same as the probability $MS_{G/T_G} < MS_{S/G/T_G}$ and is found using the fact that

$$\frac{MS_{G/T_G} \big/ \left(n\sigma^2_g + \sigma^2_{gG}\right)}{MS_{S/G/T_G} \big/ \sigma^2_{gG}} \sim F_{(g-1),g(n-1)}$$

(McGraw & Wong, 1996; Searle, 1992). Then,

$$\begin{aligned}\Pr\left\{MS_{G/T_G} < MS_{S/G/T_G}\right\} &= \Pr\{F < 1\} \\ &= \Pr\left\{\left(F_{(g-1),g(n-1)}\right) < \frac{\sigma^2_{gG}}{\left(n\sigma^2_g + \sigma^2_{gG}\right)}\right\}, \text{ or} \qquad (19) \\ &= \Pr\left\{\left(F_{(g-1),g(n-1)}\right) < \frac{1 - \rho_{ICC}}{\rho_{ICC}(n-1)+1}\right\},\end{aligned}$$

where ρ_{ICC} is the intraclass correlation.

Searle (1992) indicated negative values of $\hat{\sigma}^2_g$ are not usually a problem in balanced data if the number of groups is not too small and that having many groups is more important than having many subjects nested in the groups. Also, in the case of balanced data, the synthetic mean square of equation (15) is the same as that of the quasi-F of Myers et al. (1981) and is a linear combination of two positive terms. However, with unbalanced data, the synthetic mean square of the quasi-F statistic is not the linear combination of three positive terms. It can be shown that a_3 is always negative under the conditions of this study. Therefore, it is possible with unbalanced data to have negative estimates of σ_G^2 and σ_g^2. With unbalanced data, the denominator synthetic mean square of the quasi-F test statistic can also have zero or negative values. Increasing the degree of imbalance in the data maximizes the absolute value of a_3, and, when large absolute values of a_3 combine with values of $MS_{S/G/T_G}$ that are large in comparison to MS_{G/T_G}, the negative estimation of σ_g^2 mathematically dominates the synthetic mean square of the quasi-F test statistic.

Approximate Degrees of Freedom

Satterthwaite Two-Moment Approximation

Satterthwaite (1941) approximated the chi-square distribution of a synthetic linear combination of variance components by requiring the approximating distribution and the exact distribution to have the same first two moments. However, Satterthwaite cautioned against using the two-moment approximation when some of the coefficients in the linear combination are negative, especially when the true value of the degrees of freedom of the distribution is small. The Satterthwaite approximation for the degrees of freedom for the linear combination in equation (15) is

$$\hat{f}_2 = \frac{\left(a_1 MS_{S/T_G} + a_2 MS_{G/T_G} + a_3 MS_{S/G/T_G}\right)^2}{\frac{\left(a_1 MS_{S/T_G}\right)^2}{N_I - 1} + \frac{\left(a_2 MS_{G/T_G}\right)^2}{g - 1} + \frac{\left(a_3 MS_{S/G/T_G}\right)^2}{N_G - g}}. \tag{20}$$

For balanced data, $n_o = n$ and $a_3 = 0$; therefore, equation (20) is the sum of two positive terms and is the same expression used by Myers et al. (1981) in equation (2).

However, when data are not balanced across groups, it is possible for the denominator of the quasi-F statistic to be less than or equal to zero when the estimate of σ_g^2 is both large and negative. In these cases, as suggested by Searle (1992), it is reasonable to assume σ_g^2 is zero, and, the quasi-F statistic is replaced by the Welch t-test where

$$t_W = \frac{\left(Y_{.I} - Y_{..G}\right)}{\sqrt{\frac{MS_{S/T_I}}{N_I} + \frac{MS_{S/T_G}}{N_G}}} \text{ and}$$

$$MS_{S/T_G} = \frac{\sum_{j=1}^{g}\sum_{i=1}^{n_j}\left(Y_{ijG} - Y_{..G}\right)^2}{\left(N_G - 1\right)}$$

with two-moment degrees of freedom

$$\hat{df} = \frac{\left(\frac{MS_{S/T_I}}{N_I} + \frac{MS_{S/T_G}}{N_G}\right)}{\frac{\left(MS_{S/T_I}\right)^2}{N_I^2\left(N_I - 1\right)} + \frac{\left(MS_{S/T_G}\right)^2}{N_G^2\left(N_G - 1\right)}}$$

(Welch, 1938).

Four-Moment Approximation

Working in the context of research designs comparing the means of two treatments, each delivered to individuals, Scariano and Davenport (1986) showed that when the larger group has a smaller sampling variance for its mean, the Type I error rate can be seriously inflated. In the case of data that are balanced across groups, the ratio of the estimated sampling variance of $\overline{Y}_{..G}$ to that of $\overline{Y}_{.I}$ is $a_2 MS_{G/T_G} / a_1 MS_{S/T_I}$ and estimates

$$\frac{\left(n\sigma_g^2+\sigma_{eG}^2\right)/N_G}{\sigma_{eI}^2/N_I}$$

with corresponding degrees of freedom *(g-1)* and *(N*$(N_I\text{-}1)$. Figure 1 shows the relationship between the preceding ratio and the ratio of the respective degrees of freedom for a design that is balanced across treatment levels and groups. For simplicity, it is assumed $\sigma_G^2=\sigma_{eI}^2=1$, $N_I=N_G=24$, and the intraclass correlation is 0.20, where $\sigma_G^2=\sigma_g^2+\sigma_{eG}^2$. As can be seen, the ratio of the sampling variances increases and the ratio of corresponding degrees of freedom approaches zero as the number of groups decreases. The two-moment approach is least likely to control the Type I error rate at the nominal level when $g=2$ and $n=12$. Therefore, the Satterthwaithe, two-moment approximation used by Myers et al. (1981) might not control the Type I error rate with a small number of groups and a large number of subjects nested in each group--cases that produce large disparity between the two ratios.

Scariano and Davenport (1986) suggested a four-moment approach for approximating the chi-square distribution of the synthetic mean square of the quasi-F for balanced data. The four-moment approximate degrees of freedom are

$$\hat{f}_4=\frac{\left\{\frac{u^2}{m_1}+\frac{1}{m_2}\right\}^3}{\left\{\frac{u^3}{m_1^2}+\frac{1}{m_2^2}\right\}^2}, \tag{21}$$

where $u=\lambda_1 MS_1/\lambda_2 MS_2$, m_1 is the degrees of freedom of MS_1, and m_2 is the degrees of freedom of MS_2. In the case of balanced data, $u=\lambda_1 MS_{G/T_G}/\lambda_2 MS_{S/T_I}$,

$\lambda_1=a_2$ of equation 17, $\lambda_2=a_1$ of equation 16, m_1=$(g\text{-}1)$, and m_2= $(N_I\text{-}1)$.

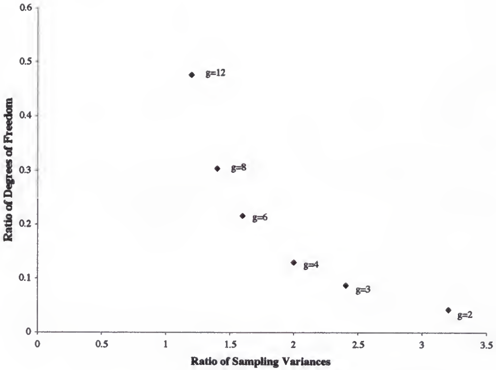

Figure 1

The Relationship of the Ratios of Degrees of Freedom and Corresponding Sampling Variances for Conditions Where $N_G = N_I = 24$, $\sigma_G^2 = \sigma_{eI}^2 = 1$, and $\rho_{ICC} = 0.2$ as *g* Increases from Two to Twelve

Scariano and Davenport also suggested as m_I increases, the four-moment quasi-F test becomes conservative and an average of the two-moment and the four-moment approximations often results in an approximation that controls the Type I error rate close to the nominal level.

Modified Four-Moment Approximation

Because the variance component terms in the synthetic error term for unbalanced data are not all positive and because of the occurrence of conditions in which the ratio of the degrees of freedom is less than one when the ratio of the corresponding sampling variances is greater than one, it is not clear the two-moment quasi-F test controls the Type I error rate at the nominal level. The four–moment approximation was developed by Scariano and Davenport (1986) for a synthetic mean square that is the sum of two positive terms. Rather than expanding the four-moment approach to three terms including one that is negative, a simpler approach that combines the two-moment and four-moment approximations was used in this study.

In order to compute the modified four-moment approximation, the group treatment level variance, σ_G^2, is first approximated using the two-moment approach. Searle (1992) showed MS_{G/T_G} and $MS_{S/G/T_G}$ are independent when data are unbalanced, and Swallow and Monahan (1984) showed the method of moments approximation works well in one-way, random effects, unbalanced ANOVA designs. The linear combination of variance components for the group treatment level data is

$$MS_{error_G} = a_2 MS_{G/T_G} + a_3 MS_{S/G/T_G} \tag{22}$$

with two-moment degrees of freedom

$$\hat{f}_2 = \frac{\left(MS_{error_G}\right)^2}{\left[\frac{(a_2 MS_{G/T_G})^2}{(g-1)} + \frac{\left(a_3 MS_{S/G/T_G}\right)^2}{(N_G - g)}\right]}. \tag{23}$$

This value of $\hat{f}_2$ along with MS_{error_G} and the estimate of the individual treatment level variance, MS_{S/T_I}, are used in the four-moment approximation of equation (21). In the modified four-moment approximation, $u = MS_{error_G} / a_1 MS_{S/T_I}$, $m_1 = \hat{f}_2$, and $m_2 = (N_I - 1)$. When the data are balanced, $a_3 = 0$, the modified four-moment approximation is the same as the four-moment approximation of equation (21). However, when $MS_{error_G} \leq 0$, the quasi-F statistic is replaced by the Welch t-test.

Average-Moment Approximation

Scariano and Davenport (1986) reported under some conditions, with completely balanced data, the four-moment quasi-F test is conservative; therefore, an arithmetic average of the two-moment and the modified four-moment approximations was also computed. In cases of unbalanced data, when $MS_{error_G} \leq 0$, data were analyzed using the Welch t-test; otherwise, the two-moment approximation and the modified four-moment approximation were arithmetically averaged resulting in a modified average moment quasi-F test.

Variables Manipulated in the Monte Carlo Study

Number of Subjects and Groups

Number of groups, *g*

In this study, I investigated the quasi-F test statistics when the number of groups was small to moderate: where the issues of statistical validity are most critical. In the three previous Monte Carlo studies, Burlingame, Kircher, & Hont (1994), Kromrey & Dickinson (1996), and Myers et al. (1981) investigated designs having 2, 3, 4, and 5 groups. In this study, I investigated 2, 3, 4, 5, and 6 groups in order to compare results with the previous studies and to determine if or when to use the two–moment, modified four-moment, or modified average-moment approximation to control the Type I error rate of the proposed quasi-F test. The same numbers of groups were investigated for both balanced and unbalanced data.

Number of subjects in groups

Myers et al. (1981) studied small groups, Burlingame, Kircher, and Honts (1994) studied groups of 8 subjects, determined by a review of psychotherapy literature as the median sized group (Burlingame, Kircher, & Taylor, 1994), and Kromrey and Dickinson (1996) studied small, medium and large groups. In this study, I examined a wide variety of group sizes in order to replicate conditions in the previous studies and to determine if or when to use two–moment, modified four-moment, or modified average-moment approximations to control the Type I error rate of the proposed quasi-F test. For the study of balanced data, group sizes of 3, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, and 30 subjects were analytically investigated and the eight sample size conditions of the Myers et al. study were simulated. For unbalanced designs, group sizes of 4, 8, 12, 16, and 20 subjects were simulated.

Number of subjects in the individual treatment level

Usually, in planned experimental designs, equal numbers of subjects are randomly assigned to both treatment levels; however, both N_I and N_G can be affected by attrition of subjects. In this study, the planned number of subjects in the individual treatment level equaled the number of subjects in the group treatment level, N_G. However, for balanced data, four combinations in which the number of subjects in the individual treatment level is equal to the number of subjects nested in each group were simulated to replicate conditions of the Myers et al. (1981) study.

The Degree of Imbalance

Imbalance in the research design can occur from naturally arising groups or from attrition of subjects in groups of planned size. Little and Rubin (1987) suggested when data from some experimental subjects are missing, the common research practice is to discard the subjects and analyze the remaining data as if it were complete. However, knowledge of any mechanism that leads to the missing values is important in the interpretation of the experimental results. Frequently the researcher has no knowledge of the mechanism that leads to attrition and assumes the mechanism is ignorable. Little and Rubin stated for sampling-based inferences, the mechanism is ignorable only when the actual observed part of the data is observed at random and the missing data are also missing at random. They referred to these conditions as missing completely at random. In this study, I assumed data are missing completely at random. Attrition rates of 15% and 25% were investigated in the Monte Carlo study. These rates were fully crossed producing a total of four attrition levels. Burlingame, Kircher, and Taylor (1994) found 18% was the median reported attrition rate in their survey of psychotherapy literature. However, in this study. I also investigated a more sizable attrition rate that produced a greater degree of imbalance.

Intraclass Correlation and Variance Components

Intraclass correlation

Following the recommendation of Swallow and Monahan (1984), the study included conditions where $\sigma_g^2 / \sigma_{\varepsilon G}^2 \leq 1$ (i.e. conditions where the method of moments estimation procedure has been shown to be adequate). Restricting $\sigma_g^2 / \sigma_{\varepsilon_G}^2 \leq 1$ also restricts the value of the intraclass correlation ρ_{ICC}. The mathematical relationship specified between $\sigma_g^2 / \sigma_{\varepsilon_G}^2$ and ρ_{ICC} results from defining $\rho_{ICC} = \sigma_g^2 / \left(\sigma_g^2 + \sigma_{\varepsilon_G}^2\right)$. Therefore, restricting $\sigma_g^2 / \sigma_{\varepsilon_G}^2 \leq 1$ simultaneously restricts $\left[\rho_{ICC} / \left(\rho_{ICC} - 1\right)\right] \leq 1$ and provides the upper boundary restriction of $\rho_{ICC} \leq 0.50$.

In the three Monte Carlo studies (Burlingame, Kircher, & Honts, 1994; Kromrey & Dickinson, 1996; Myers et al., 1981) ρ_{ICC}, or in the case of Myers et al. ρ, values of 0.00, 0.10, 0.20, 0.30, 0.40, and 0.80 were included with only 0.00 and 0.40 being common to all three studies. Intraclass correlation values of 0.00, 0.40, and 0.80 were used in the replication of the Myers et al. study and values of 0.00, 0.20, 0.40, and 0.80 were investigated in the extended analytic study of completely balanced data. In the simulation of unbalanced data, I investigated intraclass correlation values of 0.00, 0.20, and 0.40. The intraclass correlation of 0.00 reflects no group effect, 0.20 reflects a middle-sized group effect, and 0.40 reflects the single largest value manipulated in all three investigations and one near the upper boundary established by Swallow and Monahan (1984).

Variance within subjects in the individual treatment level, $\sigma_{\varepsilon_I}^2$

In all conditions $\sigma_{\varepsilon I}^2 = 1.0$.

Variance within subjects in the group treatment level $\sigma^2_{\varepsilon_G}$ and between groups σ^2_g

The variance components $\sigma^2_{\varepsilon_G}$ and σ^2_g were manipulated in order to produce three levels of intraclass correlation and three levels of the ratio of the group treatment level variance to the individual treatment level variance, $\left(\sigma^2_g + \sigma^2_{\varepsilon G}\right)/\sigma^2_{\varepsilon I}$. When $\sigma^2_g + \sigma^2_{\varepsilon G} = 1$, variances across treatment levels are homogeneous: when $\sigma^2_g + \sigma^2_{\varepsilon G} \neq 1$, variances across treatment levels are heterogeneous. When $\sigma^2_g + \sigma^2_{\varepsilon G} > 1$ or $\sigma^2_g + \sigma^2_{\varepsilon G} < 1$, the group treatment has respectively increased or decreased the overall variance of the group population. Because attrition rates effected both N_I and n_j, situations in which the larger sample size is paired with the larger variance and in which the larger sample size is paired with the smaller variance were included. Table 7 lists the variance component pairs and the respective ρ_{ICC} included in the study.

Table 7

Variance Component Pairs Manipulated in the Study

	$\rho_{ICC} = 0.00$			$\rho_{ICC} = 0.20$			$\rho_{ICC} = 0.40$			$\rho_{ICC} = 0.80$		
	$\sigma^2_G/\sigma^2_{\varepsilon I}$ 0.75	$\sigma^2_G/\sigma^2_{\varepsilon I}$ 1.00	$\sigma^2_G/\sigma^2_{\varepsilon I}$ 1.25	$\sigma^2_G/\sigma^2_{\varepsilon I}$ 0.75	$\sigma^2_G/\sigma^2_{\varepsilon I}$ 1.00	$\sigma^2_G/\sigma^2_{\varepsilon I}$ 1.25	$\sigma^2_G/\sigma^2_{\varepsilon I}$ 0.75	$\sigma^2_G/\sigma^2_{\varepsilon I}$ 1.00	$\sigma^2_G/\sigma^2_{\varepsilon I}$ 1.25	$\sigma^2_G/\sigma^2_{\varepsilon I}$ 0.75	$\sigma^2_G/\sigma^2_{\varepsilon I}$ 1.00	$\sigma^2_G/\sigma^2_{\varepsilon I}$ 1.25
σ^2_g	0.00	0.00	0.00	0.15	0.20	0.25	0.30	0.40	0.50	0.60	0.80	1.00
$\sigma^2_{\varepsilon G}$	0.75	1.00	1.25	0.60	0.80	1.00	0.45	0.60	0.75	0.15	0.20	0.25

Data Generation

The simulation in the study was carried out using the random number generation functions of the Statistical Analysis System (SAS), Release 6.12.

Individual treatment level data

Scores for simulated participants in the individual treatment level were generated using the equation

$$Y_{iI} = \mu_I + \varepsilon_{iI},$$

where μ_I was arbitrarily set at 100 and the ε_{iI}s were pseudorandom standard normal deviates generated using RANNOR. The variable Y_{iI} was set to the missing data indicator if $U_{iI} < p_I$, where p_I is the individual treatment level attrition rate and U_{iI} was a pseudorandom uniform deviate generated using RANUNI. However, N_I was always greater than or equal to two.

Group treatment level data

Scores for simulated participants in the group treatment level were generated using the equation

$$Y_{ijG} = \mu_G + \alpha_j + \varepsilon_{ijG},$$

where μ_G was arbitrarily set at 100, α_j was a pseudorandom normal deviate with mean zero and variance σ_g^2, and ε_{ijG} was a pseudorandom normal deviate with mean zero and variance $\sigma_{\varepsilon G}^2$. The variable Y_{ijG} was set to a missing value indicator if $U_{ijG} \leq p_G$, where p_G is the group treatment level attrition rate and U_{ijG} was a pseudorandom uniform deviate generated using RANUNI. However, in all cases n_j was greater than or equal to two.

Summary

In order to study the Type I error rates of the proposed quasi-F test statistics, 5472 conditions were studied. Eight sample size conditions and three intraclass correlation levels were simulated for the replication of the study of Myers et al. (1981). These three factors produced 72 conditions. Additionally, 2700 conditions of balanced data were studied analytically. The 2700 conditions resulted from a fully crossed design of 15 levels of number of subjects nested in groups, five levels of number of groups, three levels of the ratio of treatment level variances, and four levels of intraclass correlation. All conditions were analyzed using the quasi-F test statistic and the three approximations of the degrees of freedom of the denominator synthetic mean square. Furthermore, 2700 conditions of unbalanced data were simulated for a Monte Carlo study of the Type I error rates of the three quasi-F tests. These 2700 conditions resulted from crossing five levels of the number of groups, five levels of the number of subjects nested in groups, four levels of attrition, three levels of intraclass correlations, and three levels of the ratio of treatment level variances. All simulated conditions were replicated 10,000 times and the value of nominal alpha was 0.05.

CHAPTER THREE
REPLICATION OF THE STUDY OF MYERS ET AL. (1981)

This chapter discusses the replication of the study of Myers et al. The first section establishes the equivalence of the statistical models used in the original study and in the replication and provides evidence to verify the data generated for use in the replication are consistent with the statistical models. The second section presents theoretical results concerning the approximation of the degrees of freedom of the linear combination of variance components in the synthetic mean square error of the quasi-F test statistic (Scariano & Davenport, 1986). Figures 2-7 depict the approximate size of the three quasi-F tests for the groups versus individuals research design under some of the conditions of the Myers et al. study. The final section reports the results of the Monte Carlo replication of the Myers et al. study.

Verification of Data and Equivalence of Statistical Models

Statistical Models

Myers et al. (1981) modeled the group treatment level data by modeling the covariance of scores as

$$Cov\left(Y_{ijG}, Y_{i'jG}\right) = \left\{\begin{matrix} 0 \text{ when } i \neq i' \text{ and } j \neq j' \\ \sigma^2 \rho \text{ when } i \neq i' \text{ and } j = j' \\ \sigma^2 \text{ when } i = i' \text{ and } j = j' \end{matrix}\right\}$$

where ρ is the correlation between $Y_{ijG}s$ from the same group. This study modeled the group treatment level data by modeling the covariance of scores as

$$Cov\left(Y_{ijG}, Y_{i'jG}\right) = \left\{ \begin{array}{l} 0 \text{ for} i \neq i' \text{ and } j \neq j' \\ \sigma_g^2 \text{ for} i \neq i' \text{ and} j = j' \\ \sigma_g^2 + \sigma_{\varepsilon G}^2 \text{ for } i = i' \text{ and } j = j' \end{array} \right\}.$$

It can be shown when $\sigma_{\varepsilon I}^2 = 1$ and $\sigma_g^2 + \sigma_{\varepsilon G}^2 = 1,\ \rho = \rho_{ICC}$. Therefore, generating data with values of $\rho_{ICC} = 0.0$, 0.4, and 0.8 is equivalent to generating data under the conditions of Myers et al. with values of $\rho = 0.0$, 0.4, and 0.8.

Verification of Generated Data

One method to verify that generated data are consistent with the statistical model is to compare the observed number of negative estimated intraclass correlations to the probability of obtaining a negative estimated intraclass correlation. The probability is computed using the levels of the manipulated variables and equation (19). The observed and computed values are presented in Table 8 for the two-group conditions and in Table 9 for the five-group conditions. Another method to verify the generated data is to compare the intended values of the variance components with the observed values averaged over the 10,000 replications of each simulated condition. The results of this comparison are also presented in Table 8 and Table 9 for the two-group and the five-group conditions, respectively. The data simulated for this study are consistent with the statistical model. The observed variance components averaged across the 10,000 replications are consistent with intended values, and the numbers of observed negative estimates of the intraclass correlation are consistent with the probability of obtaining negative estimates.

Table 8

The Expected and Observed Occurrences of Negative Estimated Values of the Intraclass Correlation and the Expected and Observed Values of the Variance Components for Conditions Involving Two Groups

n	N_I	ρ_{ICC}	Number of Negative $\hat{\rho}_{ICC}$	Expected Number of Negative $\hat{\rho}_{ICC}$	$\bar{\hat{\sigma}}_g^2$	$\bar{\hat{\sigma}}_{eI}^2$	$\bar{\hat{\sigma}}_{eG}^2$
4	4	0.0	6480	6441	-0.0009	0.9998	1.0046
4	4	0.2	4922	4940	0.1946	1.0113	0.7989
4	4	0.4	3769	3798	0.4073	0.9926	0.5982
4	4	0.8	1809	1806	0.8156	0.9981	0.2014
4	8	0.0	6350	6441	0.0041	1.0100	.0.9976
4	8	0.2	4940	4940	0.1981	1.0009	0.7983
4	8	0.4	3798	3798	0.3981	1.0015	0.6013
4	8	0.8	1803	1806	0.7897	1.0027	0.2007
8	8	0.0	6740	6657	-0.0022	0.9977	1.0091
8	8	0.2	4223	4271	0.1980	0.9882	0.7992
8	8	0.4	2961	3029	0.3970	1.0029	0.5983
8	8	0.8	1367	1357	0.7916	1.0022	0.1986
8	16	0.0	6614	6657	0.0020	0.9950	0.9990
8	16	0.2	4230	4271	0.2009	1.0034	0.7961
8	16	0.4	3026	3029	0.4007	0.9947	0.5981
8	16	0.8	1343	1357	0.7971	1.0054	0.1994

Table 9

The Expected and Observed Occurrences of Negative Estimated Values of the Intraclass Correlation and the Expected and Observed Values of the Variance Components for Conditions Involving Five Groups

n	N_I	ρ_{ICC}	Number of Negative $\hat{\rho}_{ICC}$	Expected Number of Negative $\hat{\rho}_{ICC}$	$\bar{\hat{\sigma}}_{g}^{2}$	$\bar{\hat{\sigma}}_{\varepsilon I}^{2}$	$\bar{\hat{\sigma}}_{\varepsilon G}^{2}$
4	4	0.0	5577	5620	0.0031	0.9961	1.0004
4	4	0.2	2661	2638	0.1954	1.0038	0.7997
4	4	0.4	1124	1090	0.3939	1.0056	0.6035
4	4	0.8	54	71	0.8101	1.0059	0.1991
4	20	0.0	5688	5620	-0.0017	1.0043	1.0025
4	20	0.2	2585	2638	0.2025	1.0050	0.8027
4	20	0.4	1092	1090	0.4009	1.0046	0.6018
4	20	0.8	72	71	0.8073	0.9982	0.2006
8	8	0.0	5865	5793	-0.0019	0.9995	1.0029
8	8	0.2	1496	1464	0.1999	1.0009	0.8045
8	8	0.4	558	419	0.4073	1.0027	0.6030
8	8	0.8	17	19	0.8107	0.9896	0.1996
8	40	0.0	5779	5793	-0.0005	1.0028	0.9979
8	40	0.2	1421	1464	0.2015	1.0016	0.7996
8	40	0.4	401	419	0.4002	1.0018	0.5999
8	40	0.8	15	19	0.8009	1.0069	0.2000

The Theory of Scariano and Davenport

Scariano and Davenport (1986) provided a means for analytically determining the approximate size of a quasi-F test when comparing two treatment means at nominal alpha, α, for values of an unknown nuisance parameter $m_2\sigma_1^2 / m_1\sigma_2^2$, where σ_i^2 and m_i are the variance and degrees of freedom of population$_i$, $i = 1$ and 2. For a groups versus individuals research design that is balanced across groups, under the conditions simulated in this study, $\sigma_1^2 = \sigma_{\bar{Y}..G}^2$ which is $\frac{n\sigma_g^2 + \sigma_{\varepsilon G}^2}{N_G}$, $m_1 = (g - 1)$, $\sigma_2^2 = \sigma_{\bar{Y}.I}^2$ which is $\sigma_{\varepsilon I}^2 / N_I$, and $m_2 = (N_I - 1)$, where n is the number of subjects nested in each group. The size of the quasi-F test at nominal alpha level, α, is

$$\Pr\left[F_{quasi} > F_{(\alpha,1,\hat{f}_2)}\right] = \int_0^\infty \Pr\left[F_{quasi} > F_{[\alpha,1,\hat{f}_2]} \middle| u\right] g(u)du,$$

$$\text{where } u = \frac{a_2 MS_{G/T_G}}{a_1 MS_{S/T_I}}$$

and $\hat{f}_2$ is the Satterthwaite two-moment approximation of the degrees of freedom of the synthetic mean square error. The approximated degrees of freedom are expressed in terms of u as

$$\hat{f}_2 = \frac{(u+1)^2}{\left(\frac{u^2}{m_1} + \frac{1}{m_2}\right)}. \tag{24}$$

It can be shown that F_{quasi} is the ratio of Q, with $Q \sim F_{1,(m_1+m_2)}$, and C, a nuisance parameter, where

$$C = \frac{(1+u)(m_1+m_2)}{(1+U)\left(\frac{m_1 u}{U}+m_2\right)} \text{ and}$$

$$U = E[u] = \frac{a_2\left(n\sigma_g^2+\sigma_{eG}^2\right)}{a_1\sigma_{el}^2}.$$

An appropriate transformation of *g(u)* for numerical integration is

$$f(s) = \frac{\Gamma\left(\frac{m_1+m_2}{2}\right)}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)\cdot U\cdot(1-s)^2}\cdot\frac{\left(\frac{s}{U(1-s)}\right)^{(m_1-2)/2}}{\left(1+\frac{s}{U(1-s)}\right)^{(m_1+m_2)/2}}$$

where $s = \left[\frac{(m_1 u)}{m_2}\right]\cdot\left[1+\frac{m_1 u}{m_2}\right]^{-1}$ and $0 \le s \le 1$. The approximate size of the quasi-F test with synthetic mean square error degrees of freedom approximated using the two-moment approach ($\hat{\alpha}_{\hat{f}_2}$) is found by numerically integrating

$$\int_0^1 \Pr\left[Q > C\cdot F_{\alpha(1,\hat{f}_2)} \,\middle|\, s\right] f(s)ds. \qquad (25)$$

The approximate size of the quasi-F test using the four-moment ($\hat{\alpha}_{\hat{f}_4}$) and average-moment ($\hat{\alpha}_{\hat{f}_{ave}}$) approximations for the degrees of freedom of the synthetic mean square is found by substituting the expression for $\hat{f}_4$, equation (21), and the arithmetic average of $\hat{f}_2$ and $\hat{f}_4$, respectively, into equation (25) for $\hat{f}_2$. The numerical integration was performed using the trapezoid rule and was programmed using SAS-IML for values of *U* where $1 \le U \le 30$. For conditions when $g \ge 4$, the computation used an interval width of 0.001. When $g = 2$, a singularity occurs at $s = 0$, and, when $g = 3$, a discontinuity occurs at $s = 0$; therefore, for g ≤ 3, the limits of integration were truncated at $s = 0$ resulting in limits of integration of $.0001 \leq s \leq 1$. The truncated interval width was .0009999. This truncation kept the number of trapezoids the same for all numerical integration procedures. In several instances, the conditions of the Myers et al. study produce values of $U < 1$. Because of the properties of the F-distribution, the value of equation (25) when $U<1$ is found by interchanging m_1 and m_2 and replacing U with $1/U$. Overall, Scariano and Davenport (1986) found in all cases, $\hat{\alpha}_{\hat{f}_2} > \hat{\alpha}_{\hat{f}_4}$. However, $\hat{\alpha}_{\hat{f}_2} \rightarrow \alpha$ as $U \rightarrow \infty$ and $\hat{\alpha}_{\hat{f}_4} \rightarrow \alpha$ as $U \rightarrow \infty$. But they also reported, for small values of m_1, both $\hat{\alpha}_{\hat{f}_2}$ and $\hat{\alpha}_{\hat{f}_4}$ can deviate considerably from α as $m_2 \rightarrow \infty$. This can result in neither quasi-F test controlling the Type I error rate at nominal alpha. When $U>1$ and $m_1 > m_2$, the two-moment approximation of the degrees of freedom was shown to adequately control the size of the quasi-F test at nominal alpha.

Size of Quasi-F Tests as Determined by Numerical Integration

General conditions

The calculations of the approximate size of the quasi-F tests are based on the degrees of freedom of the mean squares and U where

$$\begin{aligned} U = E(u) &= E\left(\frac{a_2 MS_{G/T_G}}{a_1 MS_{S/T_I}}\right) \\ &= \frac{a_2 E\left(MS_{G/T_G}\right)}{a_1 E\left(MS_{S/T_I}\right)} \\ &= \frac{a_2\left(n\sigma_g^2 + \sigma_{eG}^2\right)}{a_1 \sigma_d^2}. \end{aligned}$$

Under the conditions of this study,

$$a_2 = 1/N_G,$$
$$\sigma_g^2 + \sigma_{eG}^2 = 1,$$
$$a_1 = 1/N_I,$$
$$\sigma_{eI}^2 = 1, \text{ and}$$
$$\rho_{ICC} = \sigma_g^2.$$

The results of the numerical integration are presented in Figures 2-7. The calculated approximate size of the quasi-F tests only applies to designs in which data are balanced across groups, nominal alpha = 0.05, $m_1 = (g-1)$, $m_2 = (N_1 - 1)$, and $U = a_2(n\sigma_g^2 + \sigma_{eG}^2)/a_1\sigma_{eI}^2$. However, when values of U less than one are of interest, $m_1 = (N_I - 1)$ $m_2 = (g-1)$ and $U = a_1\sigma_{eI}^2/a_2(n\sigma_g^2 + \sigma_{eG}^2)$. In all cases, the numerator degrees of freedom of the quasi-F test statistic is one.

The general conditions of the Myers et al. (1981) study are the same as listed above. Additionally, the number of subjects in the individual treatment level equaled the number of subjects in the group treatment level, $N_I = N_G = ng$, or equaled the number of subjects nested in each group, $N_I = n$. The results of the numerical integration are presented in Figures 2-7. The values of V for each Myers et al. condition, where V is consistently defined $N_I(n\sigma_g^2 + \sigma_{eG}^2)/N_G\sigma_{eI}^2$, that correspond to ρ_{ICC} values of 0.40 and 0.80 are shown on the figures. The approximate size of the three tests at ρ_{ICC} values of 0.0, 0.4, and 0.8 are also reported in Table 10 for the two-groups conditions and Table 11 for the five-groups conditions.

The Myers et al. (1981) sample size conditions

Sample size condition one--$g = 2$, $n = 4$, $N_I = 4$, $N_G = 8$--Table 10. Because $N_I \neq N_G$, the approximate size of the quasi-F when $\rho_{ICC} = 0.0$ corresponds to a value of $V<1$ and, when $\rho_{ICC} = 0.4$ and 0.8, the corresponding values of $V\geq 1$. Depiction of this condition requires two separate figures; therefore, the results are presented only in Table 10. When $\rho_{ICC} = 0$ all three quasi-F tests were conservative. The approximate size of the three tests increased as the intraclass correlation increased but $\hat{\alpha}$ remained less than nominal alpha.

Sample size condition two--$g = 2$, $n = 4$, $N_I = 8$, $N_G = 8$ -- Figure 2. All three approximations of the synthetic error term degrees of freedom resulted in liberal quasi-F tests when $\rho_{ICC} > 0.0$ with the four-moment test closest to nominal alpha. At $\rho_{ICC} = 0.0$ the two-moment test was slightly liberal, the four-moment test was slightly conservative, and the average-moment test was the closest to nominal alpha.

Sample size condition three--$g = 2$, $n = 8$, $N_I = 8$, $N_G = 16$ --Figure 3. The approximate sizes of the quasi-F tests for this condition, for values of $V\geq 1$, are shown on Figure 3. The figure is identical to Figure 2 because the values of m_1 and m_2 are the same for conditions two and three. However, the values of V corresponding to $\rho_{ICC} = 0.4$, and 0.8 are 1/2 the corresponding values of condition 2. When $\rho_{ICC} = 0.0$, $V<1$ and depiction of the results requires a separate figure. For the values of V shown on Figure 3, all three quasi-F tests were liberal. When $\rho_{ICC} = 0$, all three tests are slightly conservative with the size of the two-moment test closest to nominal alpha. The complete results of the numerical integration of this condition are reported in Table 10.

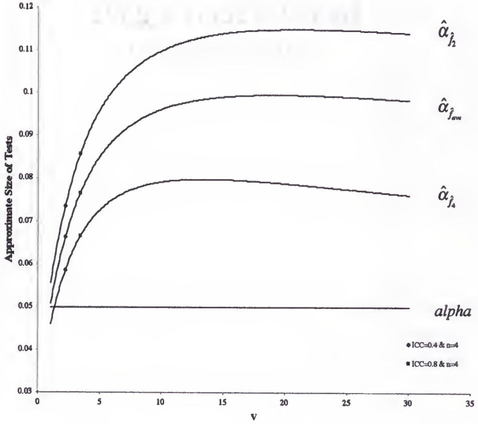

Figure 2

The Approximate Size of the Three Quasi-F Tests at m_1=1 (g-1), m_2=7 (N_L-1), and Nominal Alpha = 0.05

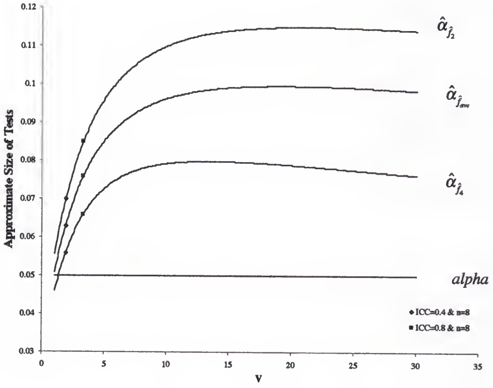

Figure 3

The Approximate Size of the Three Quasi-F Tests at m_1=1 (g-1), m_2=7 (N_l-1), and Nominal Alpha = 0.05

Table 10

The Mean Type I Error Rates of the Quasi-F Test of the Myers et al. Simulation, of Three Quasi-F Tests of the Replication of the Myers et al. Simulation, and of the Analytic Data for Conditions Involving Two Groups, Nominal Alpha = 0.05

			Simulated Data				Analytic Data		
			Myers et al.	Two-Moment	Ave-Moment	Four-Moment	Two-Moment	Ave-Moment	Four-Moment
n	N_I	ρ_{ICC}							
4	4	0.0	0.050	0.036	0.034	0.031	0.036	0.033	0.030
4	4	0.4	0.051	0.043	0.039	0.036	0.042	0.039	0.035
4	4	0.8	0.043	0.048	0.043	0.039	0.048	0.044	0.040
4	8	0.0	0.043	0.056	0.051	0.047	0.056	0.051	0.046
4	8	0.4	0.046	0.074	0.067	0.059	0.074	0.066	0.059
4	8	0.8	0.050	0.086	0.076	0.066	0.086	0.077	0.067
8	8	0.0	0.043	0.047	0.043	0.040	0.049	0.048	0.046
8	8	0.4	0.058	0.069	0.062	0.055	0.070	0.063	0.056
8	8	0.8	0.056	0.084	0.075	0.066	0.085	0.076	0.066
8	16	0.0	0.044	0.063	0.056	0.048	0.063	0.057	0.048
8	16	0.4	0.047	0.099	0.086	0.069	0.100	0.087	0.069
8	16	0.8	0.044	0.115	0.100	0.075	0.113	0.098	0.074

Sample size condition four--$g = 2$, $n = 8$, $N_I = 16$, $N_G = 16$--Figure 4. All three approximations of the degrees of freedom of the synthetic mean square resulted in liberal quasi-F tests. The approximate size of the four-moment quasi-F test was slightly conservative when $\rho_{ICC} = 0$ and became increasingly liberal as ρ_{ICC} increased. However, the size of the four-moment test showed a downward trend at values of V greater than about seven.

Sample size condition five--$g = 5$, $n = 4$, $N_I = 4$, $N_G = 20$ --Figure 5. The ratio of N_I/N_G is 1/5; therefore, the values of $\rho_{ICC} = 0.40$ and 0.80 correspond to values of V that are less than one. The approximate size of the four-moment quasi-F test was consistently conservative as predicted by Scariano and Davenport (1986). All three quasi-F tests became more conservative as ρ_{ICC} increased.

Sample size condition six--$g = 5$, $n = 4$, $N_I = 20$, $N_G = 20$-- Figure 6. The average-moment approximation of the degrees of freedom of the synthetic mean square resulted in a quasi-F test that controlled the approximate size of the test closest to the nominal level. The two-moment quasi-F test was slightly liberal and the four-moment quasi-F test was conservative. As can be seen in Figure 6, $\hat{\alpha}_{\hat{f}_2}$, $\hat{\alpha}_{\hat{f}_4}$, and $\hat{\alpha}_{\hat{f}_{ave}}$ began to slightly converge toward α as V increased.

Sample size condition seven--$g = 5$, $n = 8$, $N_I = 8$, $N_G = 40$--Table 11. The depiction of all intraclass correlation values of this condition of the Myers et al. study requires two figures; therefore, the results are reported in Table 11. The approximate size of all three tests was close to nominal alpha when $\rho_{ICC} = 0.0$. The size of the four-moment test became increasingly conservative as ρ_{ICC} increased. The size of the two-moment test remained closest to nominal alpha with $0.047 \le \hat{\alpha}_{\hat{f}_2} \le 0.049$.

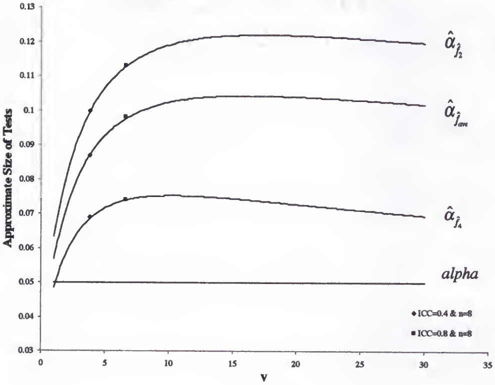

Figure 4

The Approximate Size of the Three Quasi-F Tests at m_1=1 (g-1), m_2=15 (N_f-1), and Nominal Alpha = 0.05

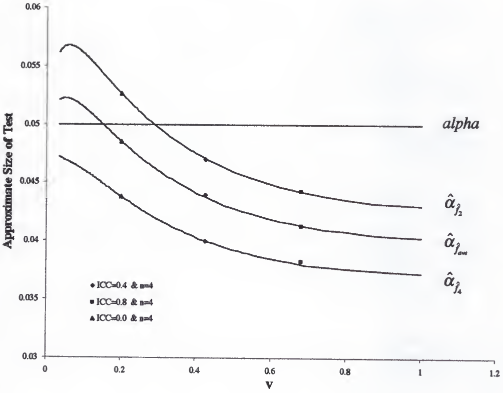

Figure 5

The Approximate Size of the Three Quasi-F Tests at m_1=4 (g-1), m_2=3 (N_1-1), and Nominal Alpha = 0.05

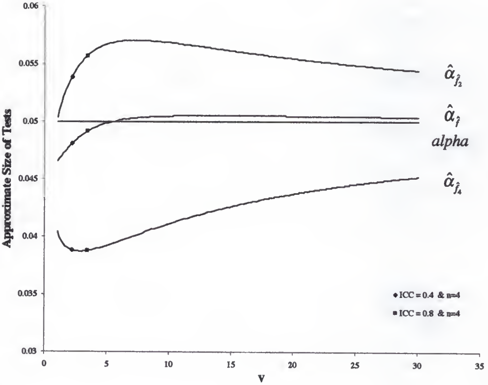

Figure 6

<u>The Approximate Size of the Three Quasi-F Tests at m_1=4 (g-1), m_2 = 19 (N_t-1) and Nominal Alpha = 0.05</u>

Table 11

The Mean Type I Error Rates of the Quasi-F Test of Myers et al. Simulation, of the Three Quasi-F Tests of the Replication of the Myers et al. Simulation, and of the Analytic Data for Conditions Involving Five Groups, Nominal Alpha = 0.05

			Simulated Data				Analytic Data		
n	N_I	ρ_{ICC}	Myers et al.	Two-Moment	Ave-Moment	Four-Moment	Two-Moment	Ave-Moment	Four-Moment
4	4	0.0	0.044	0.052	0.043	0.046	0.053	0.049	0.044
4	4	0.4	0.049	0.046	0.043	0.040	0.047	0.044	0.040
4	4	0.8	0.036	0.045	0.041	0.038	0.044	0.041	0.038
4	20	0.0	0.052	0.049	0.045	0.039	0.050	0.047	0.041
4	20	0.4	0.052	0.053	0.048	0.038	0.054	0.048	0.039
4	20	0.8	0.062	0.055	0.049	0.038	0.055	0.049	0.039
8	8	0.0	0.047	0.049	0.048	0.046	0.049	0.048	0.046
8	8	0.4	0.055	0.046	0.044	0.044	0.047	0.045	0.042
8	8	0.8	0.047	0.048	0.045	0.042	0.048	0.045	0.042
8	40	0.0	0.047	0.051	0.047	0.036	0.052	0.047	0.036
8	40	0.4	0.054	0.056	0.048	0.034	0.057	0.049	0.035
8	40	0.8	0.055	0.059	0.051	0.037	0.058	0.050	0.037

Sample size condition eight--$g = 5$, $n = 8$, $N_I = 40$, $N_G = 40$--Figure 7. The average-moment approximation of the degrees of freedom of the synthetic mean square resulted in a quasi-F test that controlled the approximate size of the test closest to the nominal level. The two-moment quasi-F test was slightly liberal and the four-moment quasi-F test was conservative. As can be seen in Figure 7, $\hat{\alpha}_{\hat{f}_2}$, $\hat{\alpha}_{\hat{f}_4}$, and $\hat{\alpha}_{\hat{f}_{ave}}$ began to slightly converge toward α as V increased.

Results of the Simulation

The results of the Myers et al. (1981) simulation, the replication of the Myers et al. simulation, and the numerical integration of the conditions of the Myers et al. study are shown for all three approaches for approximating the degrees of freedom of the synthetic error term of the quasi-F test in Tables 10 and 11. As predicted by the results of the numerical integration and Scariano and Davenport (1986), combinations of g, ρ_{ICC}, and n that produce great disparity between the ratio of the sampling variances and the ratio of the corresponding degrees of freedom resulted in quasi-F tests that did not control the approximate size of the test at nominal alpha using the two-moment approach. When g-1 is small ($g = 2$), as $(N_I - 1) \rightarrow \infty$, $\hat{\alpha}_{\hat{f}_2}$ did not converge to α under the conditions of the study. As indicated in Figures 2 and 4 $\left|\hat{\alpha}_{\hat{f}} - \alpha\right|$ increased as the number of subjects nested in the two groups increased or the value of ρ_{ICC} increased. As seen in Figures 6 and 7,the approximate size of the quasi-F tests began to slightly converge toward α as $V \rightarrow \infty$. The conservative nature of the four-moment quasi-F test when the number of groups increased is also clearly seen in Figures 6 and 7.

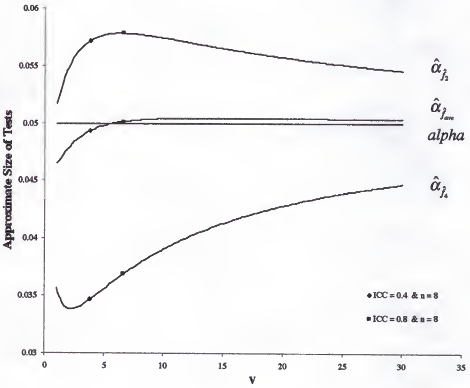

Figure 7

The Approximate Size of the Three Quasi-F Tests at m_1=4 (g-1), m_2=39 (N_f-1), and Nominal Alpha = 0.05

CHAPTER FOUR
ANALYTIC STUDY OF BALANCED DATA

This chapter presents the results of an analytic study of the quasi-F tests using the three approaches to the approximation of the degrees of freedom of the synthetic mean square error term when data are completely balanced. The conditions of the Myers et al. (1981) study were expanded to include more factors and more levels of the previously investigated factors. The same numerical methods used to calculate the approximate size of the three quasi-F tests in the replication of the Myers et al. study were used in this expanded study of balanced data. The first section presents a summary of the factors manipulated and the last section summarizes the results of the investigation of the analytic data.

Factors of the Study

I included 2700 conditions of five between-subjects factors in the expanded analytic study. In the investigation, I included three levels of the quasi-F test, resulting from $\hat{f}_2$, $\hat{f}_{ave}$, and $\hat{f}_4$, and 75 levels of sample size. The sample size conditions were five levels of the number of groups, 2, 3, 4, 5, and 6 and 15 levels of the size of the groups, starting at 3 subjects, 4 subjects, and then increasing subjects by two until reaching 30 subjects. The four levels of intraclass correlation, 0.0, 0.2, 0.4, and 0.8 were investigated. Because this study was analytic, variance component estimation was not necessary; therefore, the intraclass correlation value of 0.8 was included. Three levels of the ratio of treatment level variances (σ_G^2/σ_{el}^2) 0.75, 1.00, and 1.25 were studied. Data were obtained using the trapezoid rule to numerically integrate equation (25) at condition values of *U*. The numerical integration procedure was consistent across all conditions studied.

Results of Study

Methodology

I conducted a single subject per cell, between-subjects ANOVA to summarize the fixed effects of the five factors on the approximate size ($\hat{\alpha}$) of the quasi-F tests. The statistical model included five main, 10 two-way interaction, 10 three-way interaction, and five four-way interaction effects. All resulting 30 effects were significant, all *p*-values ≤ .0001. This model accounted for approximately 99.99% of the total variance in the approximate size of the tests. In order to interpret the results I computed a mean square component for each effect (Myers, 1979). In a study with five fixed, between-subjects factors, A, B, C, D, and E with factor levels a, b, c, d, and e, respectively, the expected value of the mean square error is

$$E(MS_{error}) = \sigma^2_{error} + \sigma^2_{S/\,ABCDE} + \theta^2_{ABCDE}.$$

The expected mean square error includes the mean square component due to interaction because there is only a single observation per cell. The variance due to the subjects component is zero because it is not possible to have more than one replication per cell. Therefore,

$$E\left(MS_{error}\right) = \sigma^2_{error} + \theta^2_{ABCDE}.$$

The mean square component of effect *j* was estimated using

$$\hat{\delta}^2_j \approx \left(\begin{array}{l} \dfrac{MS_j - MS_{error}}{a \cdot b \cdot c \cdot d \cdot e} \cdot df_j \\ 0 \text{ for } \hat{\delta}^2_j < 0 \end{array} \right),$$

where *j* is the combination of all factors in the effect and df_j are the degrees of freedom of effect *j*. It is possible that $\hat{\delta}^2_j$ underestimates the effect of combination *j* because of the confounding of σ^2_{error} and θ^2_{ABCDE} in $E\left(MS_{error}\right)$. However, the proportion of the total variance accounted for by the model suggests that θ^2_{ABCDE} is negligible. The proportion of total variance accounted for by effect *j* is

$$\hat{\omega}^2_j \approx \frac{\hat{\delta}^2_j}{\left(\sum_{j=1}^{30} \hat{\delta}^2_j\right) + MS_{error}}.$$

I considered an effect important if it accounted for more than one percent of the total variance and was significant at the $\alpha = 0.05$ level in the ANOVA table.

Results

The important effects of the study are summarized in Table 12. All of the important interactions involved the number of groups or the approximation of the degrees of freedom of the denominator synthetic mean square. The approximation of the degrees of freedom and the number of groups also contributed the most to the variance of $\hat{\alpha}$. Because of the four important two-way interactions, these will be thoroughly discussed; however, the mean approximate size of each test by the number of groups and the approach used to approximate the degrees of freedom are presented in Table 13. The mean approximate size of the two-moment quasi-F test ($\bar{\hat{\alpha}}_{\hat{f}_2}$) was consistently larger than mean approximate size the four-moment quasi-F test ($\bar{\hat{\alpha}}_{\hat{f}_4}$), and $\bar{\hat{\alpha}}_{\hat{f}_4}$ was conservative for $g \geq 3$.

Discussion of the Important Interactions

Interactions involving the denominator degrees of freedom approximation

The mean approximate size of the quasi-F tests by the number of groups--Table 13 and Figure 8. The mean approximate size of all three tests was liberal for two groups, and $\bar{\hat{\alpha}}_{\hat{f}_4}$ was conservative when $g \geq 3$. At $g = 3$, the mean approximate size of average-moment test ($\bar{\hat{\alpha}}_{\hat{f}_{ave}}$) was slightly liberal and $\bar{\hat{\alpha}}_{\hat{f}_2}$ was even more liberal. As the number of groups increased the mean approximate size of both the two-moment and average-moment quasi-F tests approached nominal alpha.

The mean approximate size of the quasi-F tests by the number of subjects nested in each group--Figure 9. Both $\bar{\hat{\alpha}}_{\hat{f}_2}$ and $\bar{\hat{\alpha}}_{\hat{f}_{ave}}$ increased as the size of the groups increased. However, $\bar{\hat{\alpha}}_{\hat{f}_4}$ was conservative and became more conservative as the number of subjects in each group increased beyond four.

Interactions involving the number of groups

The mean approximate size of the quasi-F tests by intraclass correlation--Figure 10. Averaged over all other conditions, increasing the intraclass correlation increased $\bar{\hat{\alpha}}$. The largest increase in $\bar{\hat{\alpha}}$ occurred at $g = 2$. For conditions involving three or more groups, $\bar{\hat{\alpha}}$ is conservative when $\rho_{ICC} = 0.0$ but $\bar{\hat{\alpha}}$ does not become more conservative as the number of groups increases. In conditions of four or more groups, the level of the intraclass correlation had little effect on $\bar{\hat{\alpha}}$.

Table 12

The Mean Square Components and $\hat{\omega}_j^2$ of the Important Effects of the Analytic Study of Balanced Data

Source of MS	$\hat{\omega}_j^2$
Approximation - *t*	0.273
Number of Groups -*g*	0.363
Interclass Correlation - *icc*	0.057
*t***g*	0.025
*t***n*	0.016
*g***n*	0.025
*g***icc*	0.050

Note: '*' Denotes the interaction of the main effects.

Table 13

The Mean Approximate Size of the Quasi-F Test by the Number of Groups and the Approach to the Approximation of the Synthetic Mean Square Degrees of Freedom: Nominal Alpha = 0.05

Number of Groups	$\hat{f}_2$	$\hat{f}_4$	$\hat{f}_{ave}$
2	0.0970	0.0583	0.0843
3	0.0679	0.0353	0.0574
4	0.0591	0.0338	0.0509
5	0.0555	0.0354	0.0488
6	0.0536	0.0371	0.0481

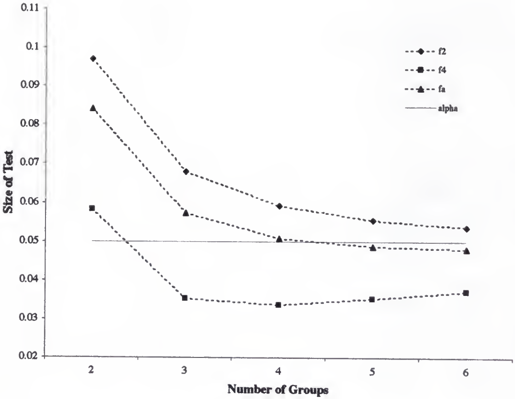

Figure 8

<u>The Effect of the Interaction of the Number of Groups and the Approach to Approximating the Degrees of Freedom of the Synthetic Mean Square on the Mean Approximate Size of the Quasi-F Tests</u>

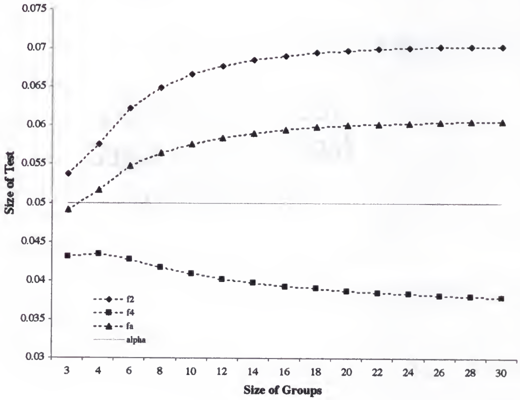

Figure 9

The Effect of the Interaction of the Number of Subjects Nested in Each Group and the Approach to Approximating the Degrees of Freedom of the Synthetic Mean Square on the Mean Approximate Size of the Quasi-F Tests

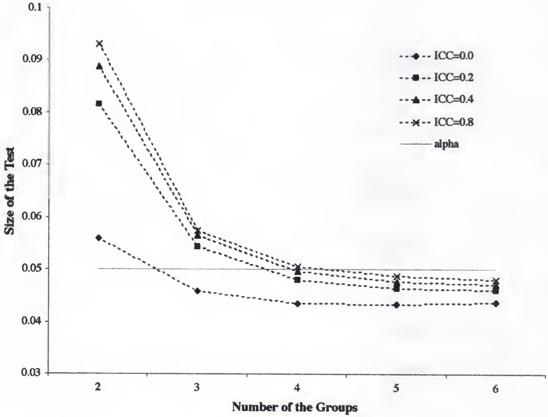

Figure 10

The Effect of the Interaction of the Number of Groups and the Level of the Intraclass Correlation on the Mean approximate Size of the Quasi-F Test: $\alpha = 0.05$

The mean approximate size of the quasi-F test by the number of subjects nested in each group --Figure 11. The mean approximate size of the quasi-F tests increased as the size of two groups increased. However, in conditions of three or more groups, increasing the number of subjects nested in each group had little effect on $\bar{\hat{\alpha}}$ of any of the tests.

Summary of the Results

Similar to the findings of Scariano and Davenport (1986), the results of the analytic study indicate $\bar{\hat{\alpha}}_{\hat{f}_2} \geq \bar{\hat{\alpha}}_{\hat{f}_{aw}} \geq \bar{\hat{\alpha}}_{\hat{f}_4}$. In cases where the ratio of the sampling variances is large and the ratio of the corresponding degrees of freedom is small, the two-moment approximation of the degrees of freedom of the synthetic mean square did not control $\bar{\hat{\alpha}}_{\hat{f}_2}$ at the nominal level. At conditions involving two groups, $\bar{\hat{\alpha}}$ of all three tests was liberal, and as either the size of the groups or the level of the intraclass correlation increased so did $\bar{\hat{\alpha}}$. Under conditions involving three groups $\bar{\hat{\alpha}}_{\hat{f}_4}$ was conservative and remained conservative as the number of groups increased. Both $\bar{\hat{\alpha}}_{\hat{f}_2}$ and $\bar{\hat{\alpha}}_{\hat{f}_{aw}}$ were slightly liberal when $g = 3$; furthermore, $\bar{\hat{\alpha}}_{\hat{f}_2}$ and $\bar{\hat{\alpha}}_{\hat{f}_{aw}}$ increased as the number of subjects nested in the three groups increased. Additionally, at $g \geq 3$ the level of the intraclass correlation had little, if any, effect on $\bar{\hat{\alpha}}_{\hat{f}_2}$ or $\bar{\hat{\alpha}}_{\hat{f}_{aw}}$. As g increased from 4 to 6 $\bar{\hat{\alpha}}_{\hat{f}_a}$ became conservative and $\bar{\hat{\alpha}}_{\hat{f}_2}$ became less liberal. The mean approximate size of both tests also increased as the size of the groups increased. However, the rate of increase was less for $\bar{\hat{\alpha}}_{\hat{f}_a}$ than $\bar{\hat{\alpha}}_{\hat{f}_a}$ and leveled off at a lower mean approximate size. Searle (1992) suggested having many groups is more important than many subjects in order to avoid negative variance component estimates. The results of this study suggest having many groups (g > 2) is more important than have many subjects nested in the groups in order to avoid an inflated Type I error rate.

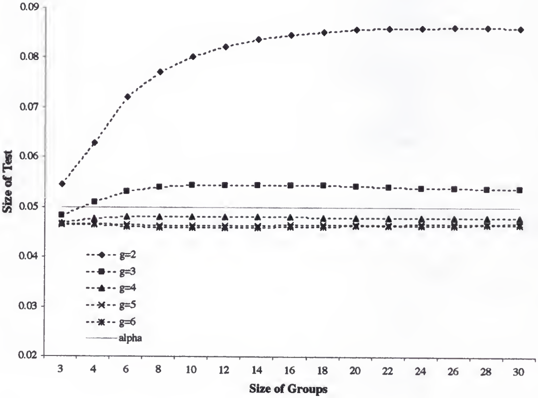

Figure 11

The Effect of the Interaction of the Number of Subjects Nested in Each Group and the Number of Groups on the Mean Approximate Size of the Quasi-F Test: $\alpha = 0.05$

CHAPTER FIVE
MONTE CARLO STUDY OF UNBALANCED DATA

This chapter presents the results of the Monte Carlo study of the quasi-F test using the three approaches for approximating the degrees of freedom of the denominator synthetic mean square of the quasi-F statistic when data are not balanced across either treatment levels or groups. Five between-subjects factors were manipulated in the study in addition to the approximation of denominator degrees of freedom. The first section of the chapter presents a summary of the factors manipulated in the simulation, and the last two sections present the results and a summary of the simulation.

Factors of the Monte Carlo Study

In the study of unbalanced data, I included 2700 conditions involving five between-subjects factors each repeated for the three levels of the one within-subjects factor. The design included three approaches to the approximation of the error term degrees of freedom as levels of the within-subjects factor. The number of groups, planned size of the groups, level of the intraclass correlation, ratio of the group to individual treatment level variances, and the rate of attrition were the five between-subjects factors. There were five levels of the number of groups, 2, 3, 4, 5, and 6; five levels of planned group size, 4, 8, 12, 16, and 20 subjects nested in the groups; three levels of intraclass correlation, 0.0, 0.2, 0.4; three levels of the ratio of group to individual treatment level variances, 0.75, 1.00, and 1.25; and four combinations of individual and group treatment level attrition rates, 0.15 and 0.15, 0.15 and 0.25, 0.25 and 0.15, and 0.25 and 0.25. Each of the conditions was replicated 10,000 times, and the Type I errors of the three tests were counted over the replications of each condition.

Results of the Monte Carlo Study

Methodology

I conducted a single subject per cell summary ANOVA to investigate the effects of the one within-subjects and five between-subjects factors. The five between-subjects factors were the number of groups, planned size of the groups, level of the intraclass correlation, ratio of the treatment level variances, and attrition rate. The one within-subjects factor was the approximation of the degrees of freedom of the synthetic mean square. The ANOVA model included 30 between-subjects effects, five main, 10 two-way, 10 three-way, and five four-way effects and 30 within-subjects effects resulting from each between-subjects effect repeated for the within-subjects effect.

In an ANOVA model with five between-subjects factors, *A*, *B*, *C*, *D*, and *E*, having levels *a*, *b*, *c*, *d*, and *e*, respectively, and a within-subjects factor *F* having *f* levels, the expected value of the model within-subjects mean square error is

$$E(MS_{ABCDEF}) = \sigma^2_{error} + \sigma^2_{SF/ABCDE} + \theta^2_{ABCDEF}$$

and the expected value of the model between-subjects mean square error is

$$E(MS_{ABCDE}) = \sigma^2_{error} + f \cdot \sigma^2_{S/ABCDE} + f \cdot \theta^2_{ABCDE}$$

(Myers, 1979). These expectations result from the omission of the six-way interaction term from the within-subjects ANOVA model and the five-way interaction term from the between-subjects ANOVA model. For a simulation study, the subject is a replication of a condition. A replication is a random event and the results cannot interact with another factor, therefore $\sigma^2_{SF/ABCDE}$ is zero. The proportion of the total variance in the approximate Type I error rate of the tests accounted for by the statistical model is 97.72% and suggests that θ^2_{ABCDEF} is negligible and is approximately equal to zero. Therefore,

$$E(MS_{ABCDEF}) \approx \sigma^2_{error}$$

and

$$E(MS_{ABCDE}) = \sigma^2_{error} + f\cdot\theta^2_{ABCDE}.$$

The variance accounted for by each effect was calculated by first setting each mean square equal to its expected value and then solving for the mean square component. For this design, the mean square component for the within-subjects effect j_{ws} is

$$\hat{\delta}^2_{j_{ws}} \approx \frac{MS_{j_{ws}} - MS_{error\,ws}}{a\cdot b\cdot c\cdot d\cdot e\cdot f}\cdot df_{j_{ws}},$$

and the mean square component for the between-subjects effect j_{bs} is

$$\hat{\delta}^2_{j_{bs}} = \frac{MS_{j_{bs}} - MS_{error_{bs}}}{a\cdot b\cdot c\cdot d\cdot e\cdot f}\cdot df_{j_{bs}}.$$

The proportion of the total variance accounted by any effect j is

$$\hat{\omega}^2_j = \frac{\hat{\delta}^2_j}{\left(\sum_{j=1}^{30}\hat{\delta}^2_{j_{ws}} + \sum_{j=1}^{30}\hat{\delta}^2_{j_{bs}} + \hat{\sigma}^2_{error} + \hat{\sigma}^2_{S/ABCDE}\right)}.$$

All negative $\hat{\delta}_j^2 s$ were set equal to zero. I considered an effect important if it accounted for more than one percent of the total variance and was significant at the $\alpha = 0.05$ level in the ANOVA table.

Results

The important effects of the Monte Carlo study are summarized in Table 14. All of the important effects of the study had ANOVA table *p*-values ≤ 0.001. The two important interactions of the between-subjects effects and the three important interactions of within-subjects effects are discussed in detail.

Between-subjects effects

The results depicted in Figure 12 indicate that, averaged over all approximations of the degrees of freedom of the synthetic mean square and the between-subjects factors other than group, $\bar{\hat{\alpha}}$ increased as the planned size of two groups increased. Increasing the planned size of three groups slightly increased $\bar{\hat{\alpha}}$, but $\bar{\hat{\alpha}}$ for groups with 12 or more planned subjects nested in the groups leveled off near the nominal alpha. Under conditions of four or more groups, $\bar{\hat{\alpha}}$ was slightly conservative and increasing the planned size of the groups had little effect on $\bar{\hat{\alpha}}$.

The results, depicted in Figure 13, suggest that, averaged over all tests and the between-subjects factors other than intraclass correlation, $\bar{\hat{\alpha}}_{icc=0.00} \leq \bar{\hat{\alpha}}_{icc=0.20} \leq \bar{\hat{\alpha}}_{icc=0.40}$; however, the effect of the level of the intraclass correlation was greater under conditions involving two groups. For conditions of four or more groups, $\bar{\hat{\alpha}}$ was slightly conservative for all levels of the intraclass correlation.

Table 14

Mean Square Components and $\hat{\omega}_j^2$ of the Important Effects of the Monte Carlo Study of Unbalanced Data

Source of MS	$\hat{\omega}_j^2$
Between Subjects Effects	
Number of Groups - *g*	0.239
Planned Size of Groups - *n*	0.024
Intraclass Correlation - *icc*	0.073
*g*n*	0.056
*g*icc*	0.079
Within-Subjects Effects	
Approximation - *t*	0.390
*t*g*	0.020
*t*n*	0.031
*t*ratio of treatment level variance*	0.017

Note: '*' Denotes the interaction of the main effects.

Within-Subjects effects

The mean approximated Type I error rates, averaged over all between-subject factors other than group, using the three approximations of the degrees of freedom of the synthetic mean square are presented in Table 15. In all cases $\overline{\hat{\alpha}}_{f_2} \geq \overline{\hat{\alpha}}_{f_{mod-ave}} \geq \overline{\hat{\alpha}}_{f_{mod-4}}$. As can be seen in Figure 14 $\overline{\hat{\alpha}}_{\hat{f}_{mod\ ave}}$ was slightly conservative under conditions of more than three groups. Only $\bar{\hat{\alpha}}_{\hat{f}_{\text{mod4}}}$, averaged across conditions involving two groups, was near nominal alpha; however, the modified four-moment approximation resulted in a conservative quasi-F tests with three or more groups. When averaged across conditions involving four groups the two-moment approximation resulted in a slightly liberal quasi-F test.

The results of the interaction of the approximation of degrees of freedom of the synthetic mean square and the planned size of the groups are depicted in Figure 15. As can be seen $\bar{\hat{\alpha}}_{\hat{f}_2}$ and $\bar{\hat{\alpha}}_{\hat{f}_{\text{mod.ave}}}$, averaged across all other conditions, increased as the planned size of the groups increased; however, $\bar{\hat{\alpha}}_{\hat{f}_{\text{mod ave}}}$ increased to only slightly liberal values, $\hat{\alpha}_{\text{mod}-ave} \le 0.056$. Again, averaged over all other conditions $\bar{\hat{\alpha}}_{\hat{f}_4}$ became increasingly more conservative as the planned size of the groups increased.

Table 15

Mean Type I Error Rates of the Quasi-F Tests by the Number of Groups and Degrees of Freedom Approximations

Number of Groups	$\hat{f}_2$	$\hat{f}_{\text{mod}-4}$	$\hat{f}_{\text{mod}-ave}$
2	0.0815	0.0497	0.0665
3	0.0635	0.0354	0.0532
4	0.0570	0.0335	0.0491
5	0.0544	0.0345	0.0479
6	0.0530	0.0360	0.0474

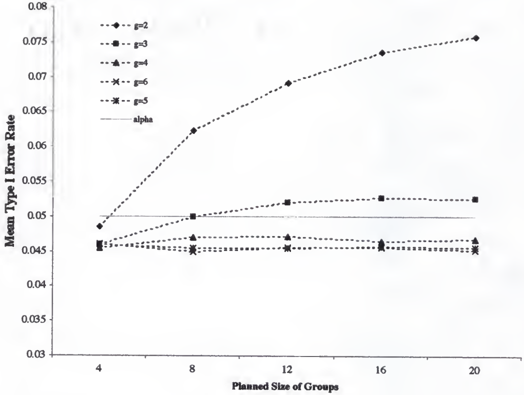

Figure 12

The Effect of the Interaction of the Number of Groups and the Planned Size of the Groups on the Mean Type I Error Rate of the Quasi-F Tests

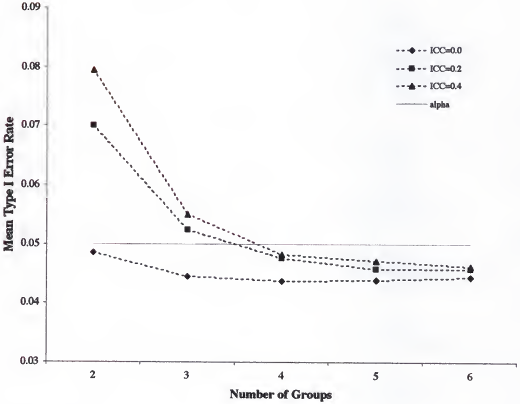

Figure 13

The Effect of the Interaction of the Number of Groups and the Level of the Intraclass Correlation on the Mean Type I Error Rate of the Quasi-F Tests

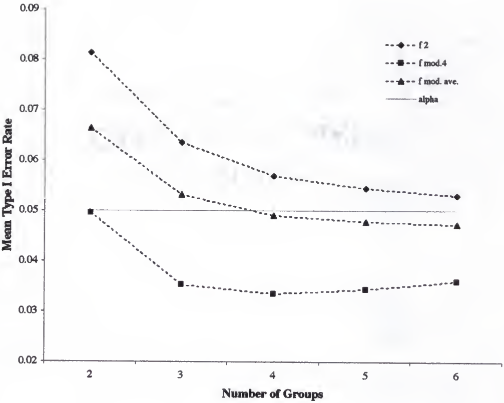

Figure 14

<u>The Effect of the Interaction of the Approximation of the Degrees of Freedom of the Synthetic Error Term and Number of Groups on the Mean Type I Error Rate of the Quasi-F Tests</u>

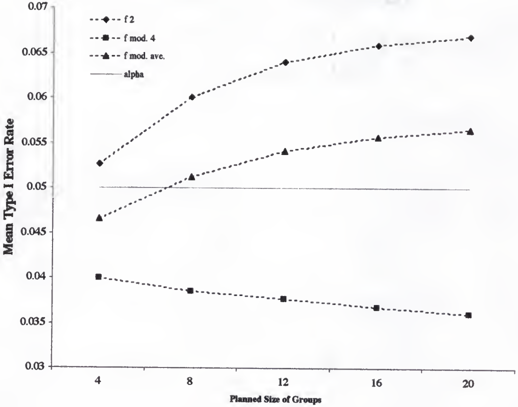

Figure 15

The Effect of the Interaction of the Approximation of the Degrees of Freedom of the Synthetic Error Term and the Planned Size of the Groups on the Mean Type I Error Rate of the Quasi-F Test

The interaction of the ratio of the treatment level variances and the approaches to the approximation of the degrees of freedom of the synthetic mean square are depicted in Figure 16. The ratio of the treatment level variances did not effect $\bar{\hat{\alpha}}_{\hat{f}_{\mathrm{mod}4}}$, averaged over all other condition. However, $\bar{\hat{\alpha}}_{\hat{f}_2}$ and $\bar{\hat{\alpha}}_{\hat{f}_{\mathrm{mod.ave}}}$ increased as the ratio of the treatment level variances increased. The increase in $\bar{\hat{\alpha}}$ was most pronounced for the two-moment quasi-F test as the ratio of treatment level variances increased from a homogeneous condition to one in which the group treatment level variance was larger.

Summary

In conditions involving two groups the two-moment approximation did not control the Type I error rate of the quasi-F test at nominal alpha. At two-group conditions $\bar{\hat{\alpha}}_{\hat{f}_2}$ increased as the planned size of the groups increased, as the level of the intraclass correlation increased, and as the ratio of the group to individual treatment level variances increased. Similarly, at the two-group conditions $\bar{\hat{\alpha}}_{\hat{f}_{\mathrm{mod.ave}}}$ increased as the planned size of the groups or the level of the intraclass correlation increased, but was less affected by an increase in the level of the ratio of the treatment level variances than the two-moment quasi-F test. The modified four-moment quasi-F test controlled the mean Type I error rate at nominal alpha in conditions involving two groups; however, $\bar{\hat{\alpha}}_{\hat{f}_4}$ increased as the planned size of the groups or the level of the intraclass correlation increased. The ratio of the treatment level variances did not affect $\bar{\hat{\alpha}}_{\hat{f}_4}$.

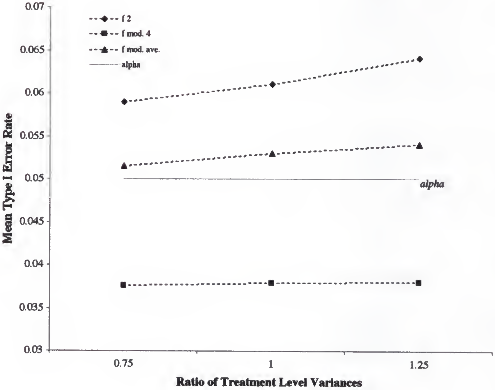

Figure 16

The Effect of the Interaction of the Approximation of the Degrees of Freedom of the Synthetic Error Term and the Ratio of the Treatment Level Variances on the Mean Type I Error Rate of the Quasi-F Tests

For conditions involving three or more groups $\bar{\hat{\alpha}}_{\hat{f}_4}$, averaged over all other conditions of the study, was very conservative. However, at conditions involving three groups both $\bar{\hat{\alpha}}_{\hat{f}_2}$ and $\bar{\hat{\alpha}}_{\hat{f}_{mod.ave}}$ were slightly liberal with the modified average-moment quasi-F test being slightly less liberal than the two-moment quasi-F test. The mean approximated Type I error rates of both tests increased as the planned size of the three groups or the level of the intraclass correlation increased; however, $\bar{\hat{\alpha}}_{\hat{f}_{mod.ave}}$ remained near nominal alpha, $\bar{\hat{\alpha}} \leq 0.055$. Again, at conditions involving three groups averaged over the other conditions of the study, increasing the ratio of the treatment level variances slightly increased $\bar{\hat{\alpha}}_{\hat{f}_{mod.ave}}$ but increased $\bar{\hat{\alpha}}_{\hat{f}_2}$ even more.

At conditions involving four or more groups the modified average-moment quasi-F test resulted in mean approximate Type I error rates at or less than nominal alpha. Increasing the level of the intraclass correlations, the planned size of the groups, or the ratio of the treatment level variances only slightly increased $\bar{\hat{\alpha}}_{\hat{f}_{mod.ave}}$. The two-moment quasi-F test was only slightly liberal, $\bar{\hat{\alpha}} \leq .057$; however, $\bar{\hat{\alpha}}_{\hat{f}_2}$ increased as the ratio of treatment level variances increased from 1.00 to 1.25, the intraclass correlation increased, or the planned size of the groups increased.

Similar to the results of the analytic study of balanced data $\bar{\hat{\alpha}}_{\hat{f}_2} \geq \bar{\hat{\alpha}}_{\hat{f}_{mod-ave}} \geq \bar{\hat{\alpha}}_{\hat{f}_{mod-4}}$. Additionally, the investigated attrition rates were not an important factor in summarizing the results of the Monte Carlo study of the Type I error rate of the quasi-F tests in designs in which data are not balanced across either treatment levels or across groups.

CHAPTER SIX
CONCLUSIONS

Myers et al. (1981) proposed a quasi-F test for comparing the means of two treatments in designs in which one treatment is delivered to individuals and the other is delivered to individuals nested in groups. The purpose of this study was to extend the work of Myers et al. in two ways. The first involved the effectiveness of two new approaches to approximating the degrees of freedom of the synthetic error term of the quasi-F test statistic in controlling the Type I error rate in data that are balanced across groups. The second involved extending the approach used by Myers et al. and the two new approaches to include data that may not be balanced across groups or treatment levels and then studying the effectiveness of the resulting quasi-F test statistics in controlling the Type I error rate. In order to provide researchers useful guidelines, the conclusions are organized by what researchers can know, the number of groups, the planned size of the groups, and whether or not the design is balanced, rather than what researchers cannot know, the intraclass correlation and the ratio of the treatment level variances. The conclusions are intended to provide researchers using the groups versus individuals research design guidelines on the use of the quasi-F tests and the three approaches for approximating the degrees of freedom of the synthetic error term. Both the factors and the levels of the factors included in the study limit the generalizibility of the results.

Conclusions

Within the limitations of the study, the following are the findings of the study:

Conclusion 1. The results of this study do not support the findings of Myers et al. (1981). Under conditions involving two groups Myers et al. reported the two-moment approach to the approximation of the degrees of freedom of the synthetic error term resulted in a quasi-F test statistic that controlled Type I error rates when $\text{Ng} = n \cdot g$ and $\rho_{ICC} > 0.0$. Both the analytic and simulated data indicated the two-moment approach resulted in a quasi-F test statistic with an inflated Type I error rate for all two-group conditions of the study. This finding is true for both balanced and unbalanced data: The Type I error rate of the two-moment quasi-F test averaged across all two-group conditions for balanced data was 0.0970 and for unbalanced data was 0.0815.

Conclusion 2. In conditions involving two groups the Type I error rate of the quasi-F test statistic using the two-moment and modified average-moment approaches for approximating the degrees of freedom of the synthetic error term for both balanced and unbalanced data increased as the planned number of subjects in the groups or the intraclass correlation increased. Additionally, for unbalanced data the Type I error rate of the two-moment and average-moment quasi-F tests increased as the ratio of the group to individual treatment level variances increased. The Type I error rate for modified four-moment test for both balanced and unbalanced data increased from four planned subjects to eight planned subjects and then remained constant (unbalanced data) or began to slightly decrease (balanced data) as the number of planned subjects increased. In conditions of four or more planned subjects the modified four-moment test was slightly liberal. The average size of the four-moment quasi-F test of balanced data, Figure 17, increased from approximately 0.05 with four subjects nested in the groups to a plateau of approximately 0.06 with more than eight subjects nested in the groups. When data are not balanced, Figure 18, the mean Type I error rate of the modified four-moment quasi-F test was slightly liberal with $\bar{\hat{\alpha}}_{\hat{f}_{mod4}} < 0.06$ for all planned two-group sizes other than four. When four subjects were planned in each group the modified four-moment quasi-F test was slightly conservative. These Type I error rates reflect attrition rates of 0.15 and 0.25 fully crossed at both treatment levels. In all cases researchers should avoid designs involving two groups. When avoidance is not possible only the four-moment or modified four-moment quasi-F tests should be used.

Conclusion 3. As depicted in Figure 17 the four-moment quasi-F test was conservative when $g \geq 3$. Also, as depicted in Figure 18 the modified four-moment quasi-F test was conservative when $g \geq 3$. Therefore, neither the four-moment nor the modified four-moment quasi-F tests are recommended for three or more groups.

Conclusion 4. In conditions involving three groups only the modified-average moment approach resulted in a quasi-F test statistic that controlled the Type I error rate near nominal alpha. For balanced conditions, Figure 17, the mean size of the average-moment quasi-F test was slightly liberal, $\bar{\hat{\alpha}}_{ave} \approx 0.06$, but the mean size did not increase as the size of the three groups increased to more than eight subjects. In cases of unbalanced data, Figure 18, the modified average-moment quasi-F test was slightly conservative with four subjects planned in each group and increased to slightly liberal with 20 subjects planned for each group. The two-moment quasi-F tests were liberal for both balanced and unbalanced conditions, $\bar{\hat{\alpha}}_{\hat{f}_2} \approx 0.07$. Therefore, for conditions involving three groups the average-moment or modified-average moment quasi-F tests are recommended.

Conclusion 5. In conditions involving four balanced groups, Figure 17, the mean size of the two-moment quasi-F test was slightly liberal and the mean size of the average-moment quasi-F test was at nominal alpha. The size of the groups had little effect on the mean size of the two-moment test for balanced groups with eight or more planned subjects; however, there was a slight increase in the mean size from four to eight planned subjects. The average-moment test was unaffected by the planned size of the groups. In conditions involving unbalanced data, Figure 18, the Type I error rate of the two-moment quasi-F test was slightly liberal and showed a slight increase from a planned four to eight subjects nested in the groups. The Type I error rate of the modified average-moment quasi-F test was slightly conservative for all planned group sizes, but was not affected by the planned size of the groups. In conditions involving four groups, researchers could use either test, depending upon the purpose of the research.

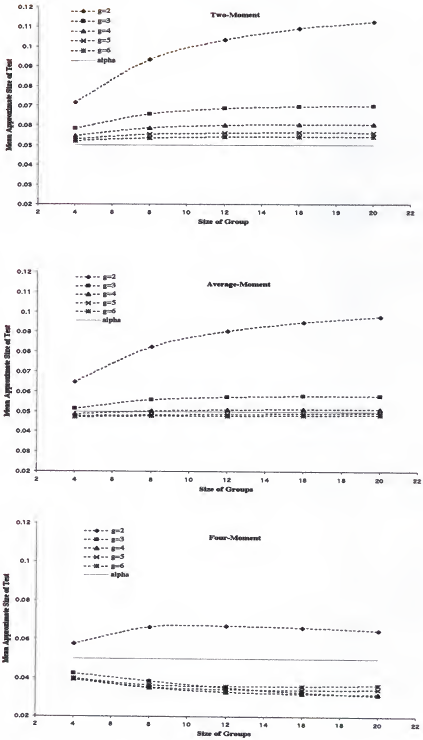

Figure 17

<u>Mean Size of the Three Quasi-F Tests by Number of Groups and Size of the Groups for Balanced Designs</u>

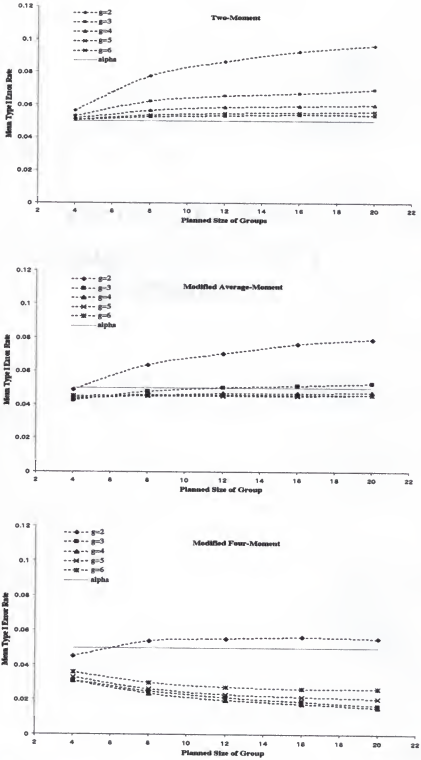

Figure 18

Mean Type I Error Rates of the Three Quasi-F Tests by Number of Groups and the Planned Size of the Groups for Unbalanced Designs

Conclusion 6. The two-moment and modified average-moment approaches resulted quasi-F tests that controlled the mean Type I error rates at nominal alpha when g > 4. In conditions involving five or six groups, Figures 17 and 18, the mean size of the two-moment quasi-F test was slightly liberal; however, the size of the group had very little effect on the mean size of the quasi-F test. The mean Type I error rate of the modified average-moment test for $g > 4$ was at nominal alpha and the planned size of the groups had no effect on the mean Type I error rate of the test. In conditions involving more than four groups, researchers could use either test, depending upon the purpose of the research.

Conclusion 7. Averaged over all conditions of the study $\overline{\hat{\alpha}}_{\hat{f}_2} \geq \overline{\hat{\alpha}}_{\hat{f}_{\text{mod ave}}} \geq \overline{\hat{\alpha}}_{\hat{f}_{\text{mod 4}}}$ and averaged over all conditions of the study by the number of groups $\overline{\hat{\alpha}}_{\hat{f}_{\text{unbalanced data}}} \leq \overline{\hat{\alpha}}_{\hat{f}_{\text{balanced data}}}$.

Conclusion 8. The levels of subject attrition rate were not an important factor in the effectiveness of the two-moment, modified average-moment, or modified four-moment approaches in controlling the Type I error rate of the quasi-F test.

Explanation of the Results

The major conclusions of this study are supported by the study of Scariano and Davenport (1986). They reported for conditions where m_1 is small (g=2) as m_2 increases toward ∞ the size of both the two-moment and four-moment quasi-F tests is larger than nominal alpha. They also suggested, under those conditions, increasing the disparity between the ratio of the treatment level sampling variances and the ratio of the corresponding degrees of freedom increases $\left|\bar{\hat{\alpha}}_{\hat{f}_2} - \alpha\right|$ and $\left|\bar{\hat{\alpha}}_{\hat{f}_4} - \alpha\right|$. Because of the design of the study, increasing the number of subjects nested in two groups also increases the number of subjects in the individual treatment level; therefore, increasing the number of subjects nested in groups increased m_2. Furthermore, when data are balanced across both treatment levels and groups, the expected value of the ratio of the sampling variances is $\left(n\sigma_g^2 + \sigma_{eG}^2\right)/\sigma_{eI}^2$. The condition of homogeneous treatment level variances and the data generation requires $\sigma_g^2 + \sigma_{eG}^2 = \sigma_{eI}^2 = 1$; therefore, increasing n or σ_g^2 increases the ratio of the treatment level sampling variance while simultaneously decreasing the corresponding ratio of the degrees of freedom. The approximated sizes of the two-moment and four-moment quasi-F tests are as Scariano and Davenport predicted.

Scariano and Davenport (1986) also reported in cases where $m_1 = 2$ ($g = 3$) the size of the two-moment quasi-F test deviated from α as m_2 increases. The mean approximate size of the two-moment quasi-F test was also larger than nominal alpha and became larger as n or ρ_{ICC} increased. The conservative nature of the four-moment quasi-F test permitted the mean approximate size of the average-moment quasi-F test to control the Type I error rate near nominal alpha. Also as predicted by Scariano and Davenport (1986) the mean approximate size of the four-moment test was conservative for g ≥ 4 and $\bar{\hat{\alpha}}_{\hat{f}_2} \geq \bar{\hat{\alpha}}_{\hat{f}_{ave}} \geq \bar{\hat{\alpha}}_{\hat{f}_4}$.

Even though Scariano and Davenport (1986) did not study unbalanced data their theory can be intuitively extended to unbalanced data. Under the conditions of this study, $n_0 \leq n$ of the corresponding balanced condition; therefore, attrition decreases the ratio of the sampling variances, and, because attrition also decreases m_2, it increases the ratio of the corresponding degrees of freedom. Following the theory of Scariano and Davenport attrition of subjects in both the individual treatment level and the subjects nested in groups should decrease the disparity between the ratio of the sampling variances and the corresponding degrees of freedom. The reduction results in smaller Type I error rate for the quasi-F tests of unbalanced data when compared to similar conditions of balanced data.

Suggestions for Future Research

The generalizibility of the results of this study is limited by the variables manipulated and the levels of those variables selected for study. The levels of the variables included a wide range of levels in order to investigate any apparent trends in the data. However, the levels of the intraclass correlation were limited by the theory of the method of moments variance component estimation procedures (Swallow & Monahan, 1984) and the levels of the attrition rate were limited by the work of Little and Rubin (1987). Among the variables not manipulated in the study are the normality of the data, variance component estimation procedures, and patterns of correlation among the scores in a group.

Micceri (1987) reported that a wide variety of psychometric distributions may not be normal and that random-effects ANOVA tests may not be robust to departures from normality, especially when condition involve unbalanced designs or small sample sizes. Departure from normality needs to be investigated in designs in which the quasi-F tests controlled the Type I error rate under conditions of normal data. Furthermore, departure from normality possibly exacerbates an already inflated Type I error rate at conditions involving two and three groups.

Burlingame, Kircher, and Honts (1994) reported that three patterns of dependency of data in treatments delivered to groups result in inflated Type I error rates when group dependency is ignored in tests of treatment effectiveness. The dependency pattern modeled in this study, as well as the studies of Burlingame, Kircher, and Honts, Kromrey and Dickinson (1996), and Myers et al. (1981), is one in which the scores of all group members are correlated with each other, called a constant correlation. Other possible patterns include one in which the score of one group member is correlated with only one other group score (serial correlation) and one in which the score of one group member is correlate with all other group scores (subgroup correlation). Burlingame, Kircher, and Honts reported the degree of Type I error rate inflation varies with both the degree and type of dependency. This study investigated only varying the degree of one type of dependency.

Additionally, the conditions involving two groups are an area for future research. Searle (1992) suggested the method of moments variance component estimation procedure might not work well with few groups and Swallow and Monahan (1984) did not investigate two-group conditions in their Monte Carlo study of variance component estimation procedures. In conditions involving two groups the probability, calculated using equation (19), of negative method of moments estimates of σ_g^2 is large even when the intraclass correlation is 0.4. Figures 19-21 show the probability of obtaining a negative estimate of σ_g^2, where $\sigma_G^2 = 1$and the intraclass correlation is 0.0, 0.2, and 0.4, respectively for two, three, four, and six groups as the size of the groups increases to thirty. Under all conditions, the probability of obtaining a negative estimate of σ_g^2 is greater than 0.50 when the intraclass correlation is actually zero. However, as suggested by Searle (1992) it is reasonable for researchers to assume σ_g^2 is actually zero in these cases.

Unfortunately, the researcher has no way of knowing σ_g^2 is actually zero. In fact, Figure 19 shows under conditions involving two groups and no group dependency the probably of obtaining a negative estimate of σ_g^2 increases to almost 70% as the size of the two groups increases. The probability of obtaining a negative estimate of σ_g^2 should be near 50% in conditions of no dependency. At conditions involving three groups and no group dependency the probability of obtaining a negative estimate of σ_g^2 increases to slightly more than 60% as the size of the groups increased. At conditions involving four and six groups and no group dependency the probability of obtaining a negative estimate of σ_g^2 is between 50% and 60% indicating σ_g^2 is still underestimated.

However, as the intraclass correlation increases to 0.2, as shown in Figure 20, the minimum probability of obtaining a negative estimate of σ_g^2 is .27, .11, .05, and .01 for 2, 3, 4, and 6 groups, respectively. When the intraclass correlation increases to 0.4, as shown in Figure 21, the minimum probability of obtaining a negative estimate of σ_g^2 is .27, .05, .01, and .00 for 2, 3, 4, and 6 groups, respectively.

Because of the probability of obtaining negative estimates of σ_g^2, Kelly and Mathew (1993) suggested the method of moments procedure always underestimates

σ_g^2 and is, therefore, unreliable. Kelly and Mathew developed several invariant quadratic estimators (IQE) of σ_g^2 that are non-negative or have a smaller probability of obtaining negative estimates than the method of moments estimation procedure. The probabilities of obtaining negative estimates of σ_g^2, shown in Figures 19-21, strongly suggest an avenue for future research is the procedure used to estimate the variance components especially in conditions of two and three groups. The inflated Type I error rates in these conditions may result from underestimating the synthetic error term rather than, or in addition to, the approach used to approximate the degrees of freedom of the synthetic error term.

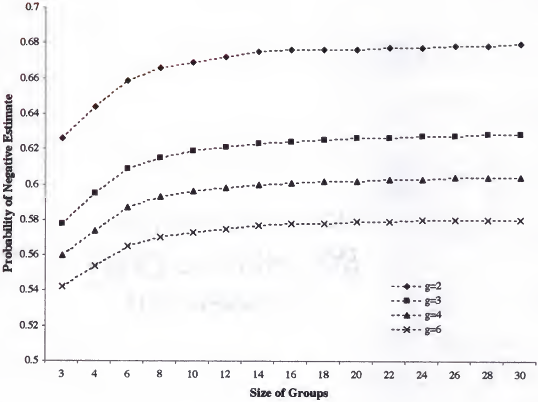

Figure 19

The Probability of Negative Estimates of σ_g^2 Under Conditions of No Group Dependency

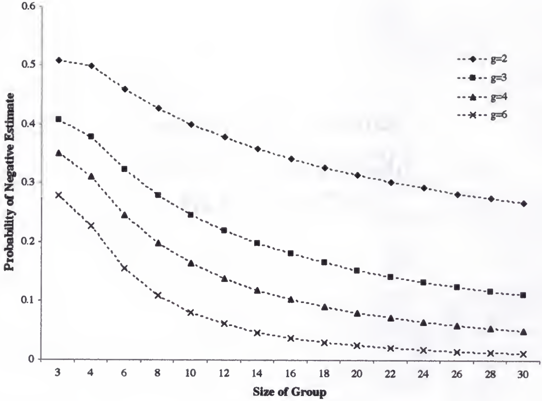

Figure 20

The Probability of Negative Estimate of σ_g^2 Under Conditions of $\sigma_g^2 = 0.20$

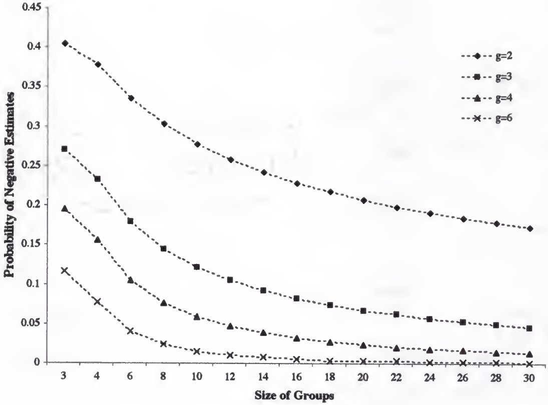

Figure 21

<u>The Probability of Negative Estimates of σ_g^2 Under Conditions of $\sigma_g^2 = 0.40$</u>

REFERENCES

Bates, G. W., Thompson, J. C., & Flanagan, C. (1999). The effectiveness of individual versus group induction of depressed mood. The Journal of Psychology, 33, 245-252.

Boling, N. C., & Robinson, D. H. (1999). Individual study, interactive multimedia, or cooperative learning: Which activity best supplements lecture-based distance education? Journal of Educational Psychology, 91, 169-174.

Burlingame, G. M., Kircher, J. C., & Honts, C. R. (1994). Analysis of variance versus bootstrap procedures for analyzing dependent observations in small group research. Small Group Research, 25, 486-501.

Burlingame, G. M., Kircher, J. C., & Taylor, S. (1994). Methodological considerations in group psychotherapy research: Past, present, and future practices. In A. Fuhriman & G. Burlingame (Eds.), Handbook of group psychotherapy and counseling: An empirical and clinical synthesis. (pp. 41-80). New York: Wiley.

Clarke, G. N. (1998). Improving the transition from basic efficacy research to effectiveness studies: Methodological issues and procedures. In A. E. Kazdin (Ed.), Methodological issues and strategies in clinical research, (2nd ed.) (pp. 541-559). New York: Wiley.

Cook, T., & Campbell, D. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston, MA: Houghton Mifflin.

Henderson, C. R. (1953). Estimation of variance and covariance components. Biometrics, 9, 226-252.

Hopkins, K. D. (1982). The unit of analysis: Group means versus individual observation. American Educational Research Journal, 19, 5-18.

Kelly, R. J., & Mathew, T. (1993). Improved estimators of variance components with smaller probability of negativity. Journal of the Royal Statistical Society. Series B (Methodological), 55, 897-911.

Kromrey, J. D., & Dickinson, W. B. (1996). Detecting unit of analysis problems in nested designs: Statistical power and Type I error rates of the F test for groups-within-treatments effects. *Educational and Psychological Measurement, 56,* 215-231.

Little, J. A., & Rubin, D. B. (1987). *Statistical analyses with missing data*. New York: Wiley.

McGraw, K.O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30-46.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156-166.

Myers, J. (1979). *Fundamentals of experimental designs*. Boston: Allyn & Bacon.

Myers, J., Dicecco, J., & Lorch, Jr., J. (1981). Group dynamics and individual performances: Pseudogroup and Quasi-F analyses. *Journal of Personality and Social Psychology, 40,* 86-98.

Milliken, G. A., & Johnson, D. E. (1992). *Analysis of messy data volume 1: Designed experiments*. Boca Raton, FL: Chapman & Hall.

Rao, C. R. (1971). Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis, 1,* 445-456.

Satterthwaite, F. W. (1941). Synthesis of variance. *Psychometrika, 6,* 309-316.

Scariano, S. M., & Davenport, J. M. (1986). A four-moment approach and other practical solutions to the Behrens-Fisher problem. *Communications in statistics: theory and methods, 15,* 1467-1504.

Searle, S. R. (1971). *Linear models*. New York: Wiley.

Searle, S. R. (1992). *Variance components*. New York: Wiley.

Snedecor, G. W., & Cochran, W. G. (1956). *Statistical methods applied to experiments in agriculture and biology* (5th Ed.). Ames, IA: Iowa State Coll. Press.

Swallow, W. H., & Monahan, J. F. (1984). Monte Carlo comparisons of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics, 26,* 47-57.

Webb, L. D. (1999). A group counseling intervention for children with Attention-Deficit Hyperactivity Disorder (Doctoral dissertation, University of Florida, 1999). Dissertation Abstracts International, 60, 3283.

Welch, B. L. (1938). On the comparison of several mean values: An alternative approach. Biometrika, 38, 330-336.

Wilcox, R. R. (1988). A new alternative to the ANOVA F and new results in James's second-order method. British Journal of Mathematical and Statistical Psychology, 41, 109-117.

BIOGRAPHICAL SKETCH

Stephanie Biller Wehry was born in Abingdon, Virginia, a small rural town in southwestern Virginia, on August 2, 1946. She attended elementary school in Warren, Ohio, and high school in Abingdon. In June 1966, she received a Bachelor of Science degree in mathematics and physics from Emory and Henry College, Emory, Virginia. In 1967 she married her college sweetheart, Allen Wehry, and began a career as the wife of a Naval Aviator. Stephanie and Allen reside in Orange Park, Florida, and have two grown children and two grandchildren.

Throughout Allen's 26-year Naval career, the family moved frequently and lived overseas. Stephanie taught high school mathematics, physics, and physical science in several states and in England. During Allen's last tour of duty, Stephanie attended the University of Memphis where she received a Master of Education in 1991 and a Master of Science, mathematics in 1993. Upon Allen's retirement from the United States Navy in 1993, the family returned to their home in Orange Park. Stephanie worked as adjunct faculty at the University of North Florida, Jacksonville, from 1993-1996 where she taught mathematics and science methods courses to pre-service elementary school teachers. She enrolled as a doctoral student in the College of Education at the University of Florida in the fall of 1996.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

James Algina

James J. Algina, Chair
Professor of Educational Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Linda M Crocker

Linda M. Crocker
Professor of Educational Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

M. David Miller

M. David Miller
Professor of Educational Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Donald H. Bernard

Donald H. Bernard
Associate Professor of Teaching and
Learning

This dissertation was submitted to the Graduate Faculty of the College of Education and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

May, 2001

M. David Miller

Chairman, Educational Psychology

Ben F. Nelms Jr.

Dean, College of Education

Dean, Graduate School